PROCEEDINGS *of the* THIRD BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY

Held at the Statistical Laboratory
University of California
26-31 December, 1954
July and August, 1955

VOLUME V

CONTRIBUTIONS TO
ECONOMETRICS, INDUSTRIAL RESEARCH, AND PSYCHOMETRY

EDITED BY JERZY NEYMAN

For further effective support of the Symposium thanks must be given the National Science Foundation, the United States Air Force Research and Development Command, the United States Army Office of Ordnance Research, and the United States Navy Office of Naval Research.

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY AND LOS ANGELES
1956

STATISTICAL INFERENCE IN FACTOR ANALYSIS

T. W. ANDERSON AND HERMAN RUBIN
COLUMBIA UNIVERSITY AND STANFORD UNIVERSITY

1. Introduction

In this paper we discuss some methods of factor analysis. The entire discussion is centered around one general probability model. We consider some mathematical problems of the model, such as whether certain kinds of observed data determine the model uniquely. We treat the statistical problems of estimation and tests of certain hypotheses. For these purposes the asymptotic distribution theory of some statistics is treated.

The primary aim of this paper is to give a unified exposition of this part of factor analysis from the viewpoint of the mathematical statistician. The literature on factor analysis is scattered; moreover, the many papers and books have been written from many different points of view. By confining ourselves to one model and by emphasizing statistical inferences for this model we hope to present a clear picture to the statistician.

The development given here is expected to point up features of model-building and statistical inference that occur in other areas where statistical theories are being developed. For example, nearly all of the problems met in factor analysis are met in latent structure analysis.

There are also some new results given in this paper. The proofs of these are mainly given in a technical Part II of the paper.

In confining ourselves to the mathematical and statistical aspects of one model, we are leaving out of consideration many important and interesting topics. We shall not consider how useful this model may be nor in what substantive areas one may expect to find data (and problems) that fit the model. We also do not consider methods based on other models. In doing this, we do not mean to imply that the model considered here is the most useful or important. It seems that this model has some usefulness and importance, it has been studied considerably, and one can give a fairly unified exposition of it.

Extensive discussion of the purposes and applications (as well as other developments) of factor analysis is given in books by psychologists (for example, Holzinger and Harmon [10], Thomson [23], Thurstone [24]). Some general discussion of statistical inference has been given in papers by Bartlett [9] and Kendall [12].

PART I. EXPOSITORY

2. The model

The model we consider is

$$X = \Lambda f + U + \mu, \tag{2.1}$$

This work was started while the authors were research associates of the Cowles Commission for Research in Economics. It has been supported in part by the Office of Naval Research.

where X, U, and μ are column vectors of p components, f is a column vector of $M(\leqq p)$ components, and Λ is a $p \times m$ matrix. We assume that U is distributed independently of f and with mean $\mathcal{E}U = 0$ and covariance matrix $\mathcal{E}UU' = \Sigma$, which is diagonal. The vector f will in some cases be treated as a random vector, and in other cases will be treated as a vector of parameters which varies from observation to observation. The vector X constitutes the observable quantities.

The most familiar interpretation of this model is in terms of mental tests. Each component of X is a score on a test or battery of tests. The corresponding component of μ is the average score of this test in the population. The components of f are the mental factors; linear combinations of these enter into the test scores. The coefficients of these linear combinations are the elements of Λ, and these are called *factor loadings*. Sometimes the elements of f are called common factors because they are common to several different tests; in the first presentation of this kind of model (Spearman [20]) f consisted of one component and was termed the general factor. A component of U is the part of the test score not "explained" by the common factors. This is considered as made up of the error of measurement in the test plus a specific factor, this specific factor having to do only with this particular test. Since in our model (with one set of observations on each individual) we cannot distinguish between these two components of the coordinate of U we shall simply term the element of U as the error of measurement.

The specification of a given component of X is similar to that in regression theory (or analysis of variance) in that it is a linear combination of other variables. Here, however, f, which plays the role of the independent variable, is not observed.

We can distinguish between two kinds of models. In one we consider the vector f to be a random vector, and in the other we consider f to be a vector of nonrandom quantities which varies from one individual to another. In the second case it would be more accurate to write $X_\alpha = \Lambda f_\alpha + U + \mu$. In the former case one sample of size N is equivalent to any other sample of size N. In the latter case, however, a set of observations $x_1, \cdots, x_N$ is not equivalent to $x_{N+1}, \cdots, x_{2N}$ because $f_1, \cdots, f_N$ will not be the same as $f_{N+1}, \cdots, f_{2N}$ and these enter as parameters. Another way of looking at this distinction is that in the latter case we have the conditional distribution of X given f. The distinction we are making is the one made in analysis of variance models (components of variance and linear hypothesis models).

When f is taken as random we shall assume $\mathcal{E}f = 0$. (Otherwise, $\mathcal{E}X = \Lambda\mathcal{E}f + \mu$, and μ can be redefined to absorb $\Lambda\mathcal{E}f$.) Let $\mathcal{E}ff' = M$. Our analysis will be made entirely in terms of first and second moments. Usually, we shall consider f and U to have normal distributions. If f is not random, then $f = f_\alpha$ for the αth individual. Then we shall assume usually

$$\frac{1}{N}\sum f_\alpha = 0 \quad \text{and} \quad \frac{1}{N}\sum f_\alpha f_\alpha' = M.$$

There is a fundamental indeterminacy in this model. Let $f = Af^*$ ($f^* = A^{-1}f$) and $\Lambda^* = \Lambda A$, where A is a nonsingular ($m \times m$) matrix. Then (2.1) can be written as

$$X = \Lambda^* f^* + U + \mu \tag{2.2}$$

where here (when f is random)

$$\mathcal{E}f^*f^{*\prime} = A^{-1}M(A^{-1})' = M^*, \tag{2.3}$$

say. If f is normal or if we only consider second-order moments, the model with Λ and f is equivalent to the model with Λ^* and f^*; that is, by observing X we cannot distinguish between these two models.

Some of the indeterminacy in the model can be eliminated by requiring that $\mathcal{E}ff' = I$, $\left(\text{or } \sum f_\alpha f_\alpha' = NI, \text{ if } f \text{ is not random}\right)$. In this case the factors are said to be *orthogonal;* if M is not diagonal, the factors are said to be *oblique.* When we assume $M = I$, then (2.3) is $A^{-1}(A^{-1})' = I$ ($I = AA'$). The indeterminacy is equivalent to multiplication by an orthogonal matrix; this is called the problem of rotation. Requiring that M be diagonal means that the components of f are independently distributed when f is assumed normal. This has an appeal to psychologists because one idea of common mental factors is (by definition) that they are independent or uncorrelated quantities.

A crucial assumption is that the components of U are uncorrelated. Our viewpoint is that the errors of observation and the specific factors are by definition uncorrelated. That is, the interrelationships of the test scores are caused by the common factors, and that is what we want to investigate. There is another point of view on factor analysis that is fundamentally quite different; that is, that the common factors are supposed to explain or account for as much of the variance of the test scores as possible. To follow this point of view, we should use a different model.

At this point we perhaps should indicate another point of view which we do not treat. That is that mental factors are positive quantities; any individual has these to some degree; each test score depends on these in a positive way. This implies that all the coefficients of Λ are nonnegative. This point of view leads to important and interesting considerations. However, in this paper we shall not consider this.

As in all problems of multivariate statistics, a geometric picture helps the intuition. We consider a p-dimensional space. The columns of Λ can be considered as m vectors in this space. They span some m-dimensional subspace; in fact, they can be considered as coordinate axes in the m-dimensional space, and f can be considered as coordinates of a point in that space referred to this particular axis-system. This subspace is called the *factor space.* Multiplying Λ on the right by a matrix corresponds to taking a new set of coordinate axes in the factor space. ($\Lambda^* = \Lambda A$ is a rotation in m-space)

3. The problems

We now list the considerations which must be made for this model. We point out that exactly the same considerations enter into other models, for example, latent structure analysis. For the sake of outlining these problems we shall assume that f is random and is normally distributed with $\mathcal{E}ff' = I$. ($M = I$)

(I) *Existence of the model.* From (2.1) we deduce that X is normally distributed with mean μ and covariance matrix

$$
\begin{aligned}
(3.1) \qquad \mathcal{E}(X-\mu)(X-\mu)' &= \mathcal{E}(\Lambda f + U)(\Lambda f + U)' \\
&= \mathcal{E}(\Lambda ff'\Lambda' + Uf'\Lambda' + \Lambda fU' + UU') \\
&= \Lambda \mathcal{E}ff'\Lambda' + \mathcal{E}UU' \\
&= \Lambda\Lambda' + \Sigma \\
&= \Psi,
\end{aligned}
$$

$\Sigma = \eta\sigma f + \mathcal{E}u + \mu$

say. Suppose we have some normal population with mean μ^* and covariance matrix Ψ^*, is there a factor analysis model that can generate this population? Essentially, this is a question whether the equation $\Psi^* = \Lambda\Lambda' + \Sigma$ can be solved, or rather, what conditions must Ψ^* satisfy so that $\Psi^* = \Lambda\Lambda' + \Sigma$ can be solved. Another way of looking at this problem leads to the formulation of what is the minimum m for which the equation can be solved.

(II) *Identification.* Suppose there is some Λ and Σ such that $\Psi^* = \Lambda\Lambda' + \Sigma$. Does this equation then have a unique solution? From the previous discussion it is clear that the above equation is also satisfied by $\Lambda\theta$, where θ is an orthogonal matrix. We can consider (i) if other restrictions are placed on Λ, is the solution unique, or (ii) are Λ and Σ determined uniquely except for multiplication of Λ on the right by an orthogonal matrix?

(III) *Determination of the structure.* Given Ψ [and suppose that (3.1) can be solved uniquely], how do we determine Λ and Σ?

We now turn to the statistical problems.

(IV) *Estimation of parameters.* A sample of N individuals is drawn, and from these observations we wish to estimate μ, Σ, and Λ. It is assumed that (3.1) can be solved uniquely and that m is known. One would like to know the properties of various estimation methods.

(V) *Test of the hypothesis that the model fits.* Here we suppose that m is given. We test the hypothesis that $\mathcal{E}(X - \mu)(X - \mu)'$ can be of the form $\Lambda\Lambda' + \Sigma$.

(VI) *Determination of the number of factors.* In many situations the number of factors m cannot be specified in advance of the statistical investigation. In these cases, the investigator wants to use as few factors as possible to "explain" the population. On what basis should he decide that he has the right number of factors?

(VII) *Other tests of hypothesis.* There are various hypotheses about the parameters, particularly about Λ, that are of interest.

(VIII) *Estimation of factor scores.* We want to make statements about the f's of our observed X's.

4. Problems of the population: Existence of the structure (I)

If f and U are normally distributed, the model postulates that the vector of p test scores X has a multivariate normal distribution with a vector of means μ and a covariance matrix Ψ which has the form

$$\Psi = \mathcal{E}(X - \mu)(X - \mu)' = \mathcal{E}(\Lambda f + U)(\Lambda f + U)' = \Lambda M \Lambda' + \Sigma \tag{4.1}$$

where Σ is diagonal and positive definite, Λ is a $p \times m$ matrix with m specified, and M is an arbitrary positive definite matrix of order m. In this case the problem of existence of the structure is the problem whether the distribution of a vector X has the above form. The question of normality will not be considered here; the vector of means μ is unrestricted and hence is of no question. The essential question is whether the covariance matrix of X has the form of (4.1); that is, given the $p \times p$ positive definite matrix Ψ, can it be expressed as $\Sigma + \Lambda M \Lambda'$ (Σ diagonal and Λ of size $p \times m$)? If f is not normal, we restrict our considerations to second-order moments, and the essential problem is the same.

As far as our present problem goes, we can assume that $M = I$, for if we are given a

matrix $\Lambda M \Lambda'$ we can write it as $\Lambda^* \Lambda^{*\prime}$ by letting $\Lambda^* = \Lambda A$, where A is a matrix such that $AA' = M$. Thus we ask if there is a Λ and Σ such that

$$\Psi = \Sigma + \Lambda\Lambda' . \tag{4.2}$$

One way of determining whether Ψ can be expressed in the desired form is to set about solving the equations

$$\psi_{ii} = \sigma_{ii} + \sum_{a=1}^{m} \lambda_{ia}^2 , \quad \text{and} \quad \psi_{ij} = \sum_{a=1}^{m} \lambda_{ia}\lambda_{ja}, \; i < j . \tag{4.3}$$

These are polynomial equations, and there are well-known methods for solving them. If there is an algebraic solution, one must ascertain that λ_{ia} are real and σ_{ii} are real and nonnegative.

The algebraic solution is laborious and gives little insight. What we want are conditions on Ψ that can be applied more directly.

A good deal of insight can be obtained by comparing the number of equations with the number of unknowns [25]. In Ψ there are $p(p+1)/2$ elements, and this is the number of equations involving the unknowns σ_{ii} and λ_{ia}. There are p elements of the diagonal Σ, and there are pm elements of Λ. However, in any solution Λ can be replaced by $\Lambda\theta$, where θ is an orthogonal $(m \times m)$ matrix, and θ has $m(m-1)/2$ independent elements; that is, in any solution, Λ can be made to satisfy $m(m-1)/2$ additional conditions. Thus the number of equations and conditions minus the number of unknowns to be determined is

$$\begin{aligned} C &= \frac{p(p+1)}{2} + \frac{m(m-1)}{2} - p - pm \\ &= \frac{(p-m)^2 - p - m}{2} . \end{aligned} \tag{4.4}$$

It can be expected that if $C \leqq 0$, then an algebraic solution to the equations is possible. If $C > 0$, one can expect that no solution is possible; in this case it appears that Ψ must satisfy some C conditions for a solution to be possible. The inequality $C \leqq 0$ can also be written

$$m \geqq \frac{2p+1-\sqrt{8p+1}}{2} = p - \frac{\sqrt{8p+1}-1}{2} . \tag{4.5}$$

Some values of p and $[2p+1-\sqrt{8p+1}]/2$ are

p	$\frac{2p+1-\sqrt{8p+1}}{2}$
1	0
3	1
5	2.3
6	3
8	4.5
9	5.2
10	6
12	7.6
13	8.4
14	9.2
15	10

This counting of equations and unknowns gives us a rough criterion of solvability; it does not, of course, lead to precise necessary and sufficient conditions for solvability. For one thing we cannot be sure that the equations are independent; another difficulty is that the solution may not be real or yield nonnegative σ_{ii}.

It is well known that a necessary and sufficient condition that a $p \times p$ matrix A can be expressed as BB', where B is $p \times m$, is that A be positive semidefinite of rank m. Thus we can state the following.

THEOREM 4.1. *A necessary and sufficient condition that Ψ be a covariance matrix of a factor analysis model with m factors is that there exist a diagonal matrix Σ^* with nonnegative elements such that $\Psi - \Sigma^*$ is positive semidefinite of rank m.*

Now the question is how we can tell whether there exists such a diagonal matrix Σ^*. It is instructive to consider the case $m = 1$. Then we can expect that Ψ has to satisfy $C = p(p-1)/2 - p$ conditions of equality as well as some inequalities. In this case Λ is a column vector and $\Lambda\Lambda'$ is a positive semidefinite matrix of rank one. The question is whether we can subtract nonnegative numbers from the diagonal elements of Ψ to give a positive definite matrix of rank one. $\Psi - \Sigma$ will be of rank one if and only if Σ can be chosen so that all second-order minors are zero. A second-order minor which does not include a diagonal element is known as a *tetrad* and has the form

$$\begin{vmatrix} \psi_{hi} & \psi_{hj} \\ \psi_{ki} & \psi_{kj} \end{vmatrix} = \psi_{hi}\psi_{kj} - \psi_{hj}\psi_{ki} \,(h, i, j, k \text{ different}) . \tag{4.6}$$

These must all be zero. A second-order minor which includes one diagonal element has the form

$$\begin{vmatrix} \psi_{ii} - \sigma_{ii} & \psi_{ij} \\ \psi_{ki} & \psi_{kj} \end{vmatrix} = (\psi_{ii} - \sigma_{ii})\,\psi_{kj} - \psi_{ij}\psi_{ki} \,(i, j, k \text{ different}) . \tag{4.7}$$

Setting this equal to zero, shows σ_{ii} must be chosen so

$$\sigma_{ii} = \psi_{ii} - \frac{\psi_{ij}\psi_{ki}}{\psi_{kj}} \qquad (\psi_{kj} \neq 0) . \tag{4.8}$$

The conditions that the solution be consistent (that is, independent of the pair j, k) are the tetrad conditions. Moreover, these conditions insure that second-order minors containing two diagonal elements are zero. It can be shown that $p(p-1)/2 - p$ of the tetrad conditions imply $\psi_{ij} = q_i q_j$ $(i \neq j)$, and this in turn implies the tetrad conditions for all i, j, k, h (all different).

If the tetrad conditions are satisfied, then $\Psi - \Sigma$ will have rank one. If this matrix is to be positive semidefinite, the diagonal elements must be nonnegative; that is, $\psi_{ki}\psi_{ij}/\psi_{kj} \geqq 0$. If Σ is to be positive semidefinite, $\sigma_{ii} \geqq 0$.

THEOREM 4.2. *A necessary and sufficient condition that Ψ be a covariance matrix of a factor analysis model with one factor is that $p(p-1)/2 - p$ independent tetrad conditions are satisfied and*

$$0 \leqq \frac{\psi_{ki}\psi_{ij}}{\psi_{kj}} \leqq \psi_{ii} \tag{4.9}$$

for one pair $(j \neq k)$ for each i.

Another way of expressing the condition $\psi_{ki}\psi_{ij}/\psi_{kj} \geqq 0$ is to ask whether one can multiply some rows and corresponding columns by -1 to obtain a matrix with all nonnegative elements.

The case of one factor is of particular interest. In fact, the original theory of Spearman was given for one "general" factor.

A similar analysis can be made for the case $m = 2$. However, the conditions become more complicated (see [26]).

In section 8 we shall consider the question of determining from the sample whether a factor analysis model with a given number of factors is adequate to "explain" the situation. The study of the problem of solvability in the population is of importance for the insight it gives into the model and for suggestions of how to use the sample to ascertain whether the model is suitable.

Albert [1] has given a theorem that leads to a direct procedure for determining whether $\Psi - \Sigma$ is of rank m. (The procedure does not verify whether $\Psi - \Sigma$ is positive definite.) Suppose that m is the maximum rank of the submatrices of Ψ that do not include diagonal elements. Then the rows and columns of Ψ can be numbered so

$$\Psi = \begin{pmatrix} \Psi_{11} & \Psi_{12} & \Psi_{13} \\ \Psi_{21} & \Psi_{22} & \Psi_{23} \\ \Psi_{31} & \Psi_{32} & \Psi_{33} \end{pmatrix} \tag{4.10}$$

where Ψ_{11}, $\Psi_{12} = \Psi_{21}'$, and Ψ_{22} are square submatrices of order m and Ψ_{12} is nonsingular. Then $\Psi - \Sigma$ is of rank m if

$$\begin{aligned} &\Psi_{12} = (\Psi_{11} - \Sigma_1)\Psi_{21}^{-1}(\Psi_{22} - \Sigma_2)\,, \Psi_{13} = (\Psi_{11} - \Sigma_1)\Psi_{21}^{-1}\Psi_{23} \\ &\Psi_{32} = \Psi_{31}\Psi_{21}^{-1}(\Psi_{22} - \Sigma_2)\,, \Psi_{33} - \Sigma_3 = \Psi_{31}\Psi_{21}^{-1}\Psi_{23}\,. \end{aligned} \tag{4.11}$$

Albert [2] has further shown that if Ψ_{31} and Ψ_{32} are also of rank m, then there is a uniquely determined Σ such that $\Psi - \Sigma$ is of rank m.

5. Problems of the population: Identification (II)

Here we assume that there is at least one solution to $\Psi = \Sigma + \Lambda\Lambda'$, and we ask whether there is more than one solution. More precisely, we assume that there is at least one solution satisfying some conditions and ask whether there is more than one solution satisfying these conditions. Since any solution Σ, Λ can be replaced by Σ, $\Lambda\theta$, where θ is orthogonal, it is clear that if we are to have a unique solution, some additional conditions must be put on Λ and Σ.

We can distinguish between two kinds of sets of restrictions. A set of one kind will not affect $\Lambda\Lambda'$, while a set of the other kind may limit $\Lambda\Lambda'$. A set of restrictions of the first kind is essentially a mathematical convenience, for any solution Σ, Λ gives a whole class of solutions Σ, $\Lambda\theta$ and a set of restrictions of the first kind simply picks out of $\Lambda\theta$ a representative solution. It is fairly clear how we can go from the class of solutions to the representative one and how we can generate the class from the representative solution.

In section 4 we noted that there are $p(p+1)/2$ elements of Ψ, p elements of Σ, pm elements of Λ and $m(m-1)/2$ independent elements of θ. We can expect that $m(m-1)/2$ restrictions will be needed to eliminate the indeterminacy due to θ. If $C = \frac{1}{2}[(p-m)^2 - p - m]$ is nonnegative we can then expect identification. If C is negative, we can expect that $-C$ additional restrictions are necessary for identification; in this case there should be in all $-C + m(m-1)/2 = p + pm - p(p+1)/2$.

This counting of equations is, of course, inadequate for making precise statements about identification. We shall now investigate the problem more adequately. It is possible

to consider conditions on Ψ that imply identification (that is, unique solvability) just as in the previous section we considered conditions on Ψ for solvability. However, it is more convenient to consider conditions involving Σ and Λ for the one assumed solution. We shall first consider conditions assuming that Σ and $\Lambda\Lambda'$ are determined uniquely.

LEMMA 5.1. *If*

$$LL' = \Lambda\Lambda' \tag{5.1}$$

where Λ and L are $p \times m$ and Λ is of rank m, then $L = \Lambda\theta$, where θ is orthogonal.

PROOF. The lemma is well known, but we give a proof for the sake of completeness; methods for finding L subject to certain restrictions are given in section 6. Since Λ is of rank m, $\Lambda\Lambda'$ is of rank m and L must be of rank m. Multiply (5.1) on the right by $L(L'L)^{-1}$ to obtain $L = \Lambda B$, where $B = \Lambda' L(L'L)^{-1}$. Multiplication on the left by $(L'L)^{-1}L'$ shows $I = B'B$. Q.E.D.

THEOREM 5.1. *A sufficient condition for identification of Σ and Λ up to multiplication on the right by an orthogonal matrix is that if any row of Λ is deleted there remain two disjoint submatrices of rank m.*

PROOF. Let $\Psi = \Sigma + \Lambda\Lambda'$. To prove the theorem we shall now show that if $\Psi = S + LL'$, where S is diagonal and L is $p \times m$, then $S = \Sigma$ and $LL' = \Lambda\Lambda'$. Since the off-diagonal elements of $\Lambda\Lambda'$ and of LL' are the corresponding off-diagonal elements of Ψ, we only have to show that the diagonal elements of LL' are equal to the diagonal elements of $\Lambda\Lambda'$.

The condition implies that $2m + 1 \leqq p$. Let

$$\Lambda = \begin{pmatrix} \Lambda_1 \\ \lambda_{m+1} \\ \Lambda_2 \\ \Lambda_3 \end{pmatrix} \qquad L = \begin{pmatrix} L_1 \\ l_{m+1} \\ L_2 \\ L_3 \end{pmatrix} \tag{5.2}$$

where Λ_1 and Λ_2 are nonsingular, and λ_{m+1} is the $(m+1)$st row; L is partitioned in submatrices of the same number of rows. Then

$$\Lambda\Lambda' = \begin{pmatrix} \Lambda_1\Lambda_1' & \Lambda_1\lambda_{m+1}' & \Lambda_1\Lambda_2' & \Lambda_1\Lambda_3' \\ \lambda_{m+1}\Lambda_1' & \lambda_{m+1}\lambda_{m+1}' & \lambda_{m+1}\Lambda_2' & \lambda_{m+1}\Lambda_3' \\ \Lambda_2\Lambda_1' & \Lambda_2\lambda_{m+1}' & \Lambda_2\Lambda_2' & \Lambda_2\Lambda_3' \\ \Lambda_3\Lambda_1' & \Lambda_3\lambda_{m+1}' & \Lambda_3\Lambda_2' & \Lambda_3\Lambda_3' \end{pmatrix}, \tag{5.3}$$

and LL' has the same form. Since $\Lambda_1\lambda_{m+1}'$, $\lambda_{m+1}\Lambda_2'$ and $\Lambda_1\Lambda_2'$ are off-diagonal, $\Lambda_1\lambda_{m+1}' = L_1 l_{m+1}'$, $\lambda_{m+1}\Lambda_2' = l_{m+1}L_2'$, and $\Lambda_1\Lambda_2' = L_1L_2'$, which is nonsingular (since Λ_1 and Λ_2 are nonsingular). Since LL' is of rank m

$$\begin{aligned} 0 = \begin{vmatrix} L_1 l_{m+1}' & L_1L_2' \\ l_{m+1}l_{m+1}' & l_{m+1}L_2' \end{vmatrix} &= \begin{vmatrix} \Lambda_1\lambda_{m+1}' & \Lambda_1\Lambda_2' \\ l_{m+1}l_{m+1}' & \lambda_{m+1}\Lambda_2' \end{vmatrix} \\ &= (-1)^m l_{m+1}l_{m+1}' |\Lambda_1\Lambda_2'| + f(\Lambda)\,. \end{aligned} \tag{5.4}$$

Similarly, $0 = (-1)^m\lambda_{m+1}\lambda_{m+1}'|\Lambda_1\Lambda_2'| + f(\Lambda)$. Since $|\Lambda_1\Lambda_2'| \neq 0$, $l_{m+1}l_{m+1}' = \lambda_{m+1}\lambda_{m+1}'$. In the same fashion, we show that the other diagonal elements of LL' are equal to those of $\Lambda\Lambda'$. This proof is patterned after Albert [1].

We can give a geometric interpretation of this condition. The columns of Λ are vec-

tors in p-space; the columns of Λ after a row is deleted are the projections of the vectors on the space of $p - 1$ coordinate axes. We require that the projection of these vectors on two different spaces of m coordinate axes span the two spaces.

It is fairly clear that the condition is unnecessarily strong in general. After one communality, that is, diagonal element of LL', is determined, it can be used in determining another. Moreover, the condition that $2m + 1 \leqq p$ is much stronger than that $C \geqq 0$. Wilson and Worcester [27] have given an example of $p = 6$ and $m = 3$ where one and only one solution exists.

We now give some theorems that include necessary conditions for identification. It will be assumed now that Σ is positive definite.

THEOREM 5.2. *Let $C_m(\Lambda)$ be a condition on Λ that is necessary for identification. Then $C_m(\Lambda\theta)$ for any orthogonal θ is also a necessary condition for identification.*

PROOF. If $C_m(\Lambda)$ is not true, there is an S and an L such that

$$\Lambda\Lambda' + \Sigma = LL' + S \tag{5.5}$$

and $\Lambda\Lambda' \neq LL'$. If $C_m(\Lambda\theta)$ is not true, then there is an S^* and L^* such that

$$(\Lambda\theta)(\Lambda\theta)' + \Sigma = L^*L^{*\prime} + S^* \tag{5.6}$$

and $\Lambda\theta(\Lambda\theta)' \neq L^*L^{*\prime}$, but the equation implies $\Lambda\Lambda' + \Sigma = L^*L^{*\prime} + S^*$ and $\Lambda\Lambda' \neq L^*L^{*\prime}$.

THEOREM 5.3. *Let $C_m(\Lambda)$ be a condition on Λ that is necessary for identification. Let Λ^* be a submatrix formed by taking m^* columns of Λ. Then $C_m^*(\Lambda^*)$ is a necessary condition for identification.*

PROOF. Let the columns of Λ be arranged so that $\Lambda = (\Lambda^*\Lambda^{**})$. If $C_m^*(\Lambda^*)$ is not true, there is an S and an L^* such that

$$\Lambda^*\Lambda^{*\prime} + \Sigma = L^*L^{*\prime} + S \tag{5.7}$$

and $\Lambda^*\Lambda^{*\prime} \neq L^*L^{*\prime}$. Then (5.5) is satisfied for $L = (L^*\Lambda^{**})$ and $\Lambda\Lambda' = \Lambda^*\Lambda^{*\prime} + \Lambda^{**}\Lambda^{**\prime} \neq L^*L^{*\prime} + \Lambda^{**}\Lambda^{**\prime} = LL'$.

THEOREM 5.4. *Let $C_{m,p}(\Lambda)$ be a condition on Λ that is necessary for identification. Let Λ^* be the matrix derived from Λ by deleting the rows that have only zero elements. Then $C_{m,p}{}^*(\Lambda^*)$ is a necessary condition for identification.*

PROOF. Let the rows be numbered so

$$\Lambda = \begin{pmatrix} \Lambda^* \\ 0 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma^* & 0 \\ 0 & \Sigma^{**} \end{pmatrix}, \qquad \Psi = \begin{pmatrix} \Psi^* & 0 \\ 0 & \Psi^{**} \end{pmatrix}. \tag{5.8}$$

Then $\Psi = \Lambda\Lambda' + \Sigma$ becomes

$$\Psi^* = \Lambda^*\Lambda^{*\prime} + \Sigma^*, \tag{5.9}$$

$$\Psi^{**} = \Sigma^{**}, \tag{5.10}$$

and only the first involves Λ^* and Σ^*.

LEMMA 5.2. *If $p = 2$ and $m = 1$, $\Lambda\Lambda'$ and Σ are not identified.*

PROOF. In this case

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11} + \lambda_{11}^2 & \lambda_{11}\lambda_{21} \\ \lambda_{21}\lambda_{11} & \sigma_{22} + \lambda_{21}^2 \end{pmatrix}. \tag{5.11}$$

If one component of Λ is 0, say λ_{21}, then $\psi_{12} = \psi_{21} = 0$ and $\psi_{22} = \sigma_{22}$. Then we can take $s_{22} = \sigma_{22}$, $l_{21} = 0$, and s_{11} and l_{11} as any numbers satisfying $s_{11} + l_{11}^2 = \psi_{11}$ ($\geqq 0$). If $\lambda_{21} \neq 0$, then $\psi_{12} \neq 0$. Let l_{21} be any number so that $\psi_{12}^2/\psi_{11} < l_{21}^2 < \psi_{22}$. Then we take $s_{22} = \psi_{22} - l_{21}^2$, $l_{11} = \psi_{12}/l_{21}$, and $s_{11} = \psi_{11} - l_{11}^2 = \psi_{11} - \psi_{21}^2/l_{21}^2$.

THEOREM 5.5. *A necessary and sufficient condition for identification if $m = 1$ is that at least three factor loadings be nonzero.*

PROOF. The necessity follows from lemma 5.2 and theorem 5.4; the sufficiency is a special case of theorem 5.1.

THEOREM 5.6. *A necessary condition for identification is that each column of ΛA has at least three nonzero elements for every nonsingular A.*

PROOF. For $A = I$, the result follows from theorems 5.3 and 5.5. Then theorem 5.2 implies the result for A being orthogonal. If A is not orthogonal, suppose ΛA has less than three nonzero elements in the νth column. Then the same will be true for an orthogonal matrix with νth column proportional to the νth column of A.

LEMMA 5.3. *If $p = 4$ and $m = 2$, $\Lambda\Lambda'$ and Σ are not identified.*

PROOF. Let the rows of Λ be numbered so there is a nonzero element in the first row (theorem 5.6). We can multiply Λ on the right by an orthogonal matrix so Λ has the form

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_1^* & \lambda_2^* \end{pmatrix}, \tag{5.12}$$

where $\lambda_{11} \neq 0$. All components of λ_2^* are nonzero by theorem 5.6. We shall now find L of the form of Λ so

$$\begin{pmatrix} \sigma_{11} + \lambda_{11}^2 & \lambda_{11}\lambda_1^{*\prime} \\ \lambda_{11}\lambda_1^* & \Sigma^* + \lambda_1^*\lambda_1^{*\prime} + \lambda_2^*\lambda_2^{*\prime} \end{pmatrix} = \begin{pmatrix} s_{11} + l_{11}^2 & l_{11}l_1^{*\prime} \\ l_{11}l_1^* & S^* + l_1^*l_1^{*\prime} + l_2^*l_2^{*\prime} \end{pmatrix}. \tag{5.13}$$

Let $l_{11} = k\lambda_{11}$, where $k > 1$ and $s_{11} = \sigma_{11} + \lambda_{11}^2 - l_{11}^2 = \sigma_{11} + (1 - k^2)\lambda_{11}^2 > 0$. Let $l_1^* = (1/k)\lambda_1^*$. Then

$$\begin{aligned} S^* + l_2^*l_2^{*\prime} &= \Sigma^* + \lambda_1^*\lambda_1^{*\prime} + \lambda_2^*\lambda_2^{*\prime} - l_1^*l_1^{*\prime} \\ &= \Sigma^* + \left(1 - \frac{1}{k^2}\right)\lambda_1^*\lambda_1^{*\prime} + \lambda_2^*\lambda_2^{*\prime} \end{aligned} \tag{5.14}$$

is positive definite. If $1 - 1/k^2$ is taken small enough, the nondiagonal elements of the right-hand side of (5.14) have the same signs as the corresponding elements of $\lambda_2^*\lambda_2^{*\prime}$. By theorem 4.2 there is a solution of (5.14) for S and l_2^*.

THEOREM 5.7. *A necessary and sufficient condition for identification if $m = 2$ is that if any row of Λ is deleted, the remaining rows of Λ can be arranged to form two disjoint matrices of rank 2.*

PROOF. The sufficiency is a special case of theorem 5.1. To prove the necessity suppose that if we delete the first row of Λ there are not two remaining disjoint matrices of rank 2. Let the rows of Λ be arranged so Λ can be partitioned as

$$\Lambda = \begin{pmatrix} \lambda_1 \\ \Lambda_2 \\ \Lambda_3 \end{pmatrix} \tag{5.15}$$

where Λ_2 is 2×2 and of rank at most 2, and Λ_3 is of rank 1. Since Λ_3 is of rank 1, there

is an orthogonal matrix θ such that $\Lambda_3\theta = (\nu\ 0)$, where ν and 0 are vectors of $p-3$ components. Let

$$\Lambda\theta = \Lambda^* = \begin{pmatrix} \lambda_{11}^* & \lambda_{12}^* \\ \lambda_{21}^* & \lambda_{22}^* \\ \lambda_{31}^* & \lambda_{32}^* \\ \nu & 0 \end{pmatrix}. \tag{5.16}$$

By theorem 5.6, $\lambda_{22}^* \neq 0 \neq \lambda_{32}^*$. After deleting the first row of Λ^*, we can get two submatrices of rank 2 only in the form

$$\begin{pmatrix} \lambda_{21}^* & \lambda_{22}^* \\ \nu_i & 0 \end{pmatrix}, \quad \begin{pmatrix} \lambda_{31}^* & \lambda_{32}^* \\ \nu_j & 0 \end{pmatrix}. \tag{5.17}$$

The assumption that there are not two such matrices of rank 2 implies that $\nu_i = 0$ except for at most one index i. Then theorem 5.4 and lemma 5.3 imply Λ^* (and Λ) is not identified.

THEOREM 5.8. *A necessary condition for identification is that for each pair of columns of ΛA and for every nonsingular A when a row is deleted, the remaining rows of this two-column matrix can be arranged to form two disjoint submatrices of rank* 2.

PROOF. This follows from theorems 5.2, 5.3, and 5.7.

Now let us consider restrictions that eliminate the indeterminacy of rotation. We might note in passing that we consider Λ and Λ^* as equivalent if each column of Λ^* is obtained by multiplying the column of Λ by ± 1, for replacing a column of Λ by its negative is only equivalent to replacing a factor score by its negative. Each of the following set of restrictions is convenient for a particular method of solving $C = \Lambda\Lambda'$ for Λ (section 6) and for a method of estimation.

(*a*) *Triangular matrix of* 0's. This condition is that

$$\Lambda = \begin{pmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda_{m1} & \lambda_{m2} & \lambda_{m3} & \cdots & \lambda_{mm} \\ \vdots & \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \lambda_{p3} & \cdots & \lambda_{pm} \end{pmatrix}; \tag{5.18}$$

that is, that the upper square matrix is triangular. If we think of a row of Λ as a vector in m-space, the condition is that the first row coincide with the first coordinate axis, the second row lie in the plane determined by the first two coordinate axes, etc.

(*b*) *General triangularity condition.* Let B be a given $p \times m$ matrix (of rank m). Here we require that

$$B'\Lambda = \begin{pmatrix} x & 0 & 0 & \cdots & 0 \\ x & x & 0 & \cdots & 0 \\ x & x & x & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ x & x & x & \cdots & x \end{pmatrix}, \tag{5.19}$$

where x indicates an element not specified zero. It is seen that if $B' = (I\ 0)$, then we obtain condition (a).

(c) *Diagonality of* $\Lambda'\Lambda$. Here we require that $\Lambda'\Lambda$ be diagonal and that the diagonal elements of $\Lambda'\Lambda$ be different and arranged in descending order. Given a positive definite matrix A, there is a uniquely determined orthogonal matrix θ (except for multiplication of columns by -1) such that $\theta' A\theta$ is diagonal with diagonal elements arranged in descending order assuming that the diagonal elements (which are the characteristic roots of A) are different. If A is already in this diagonal form, $\theta = I$.

(d) *Diagonality of* $\Lambda'\Sigma^{-1}\Lambda$. Here we require that $\Lambda'\Sigma^{-1}\Lambda$ be diagonal and that the diagonal elements be different and arranged in descending order. Rao [17] has related this condition to canonical correlation analysis.

The conditions given above are more or less arbitrary ways of determining the factor loadings uniquely. They do not correspond to any theoretical considerations of psychology; there is no inherent meaning on them. We shall now consider two types of restrictions on Λ which may have intrinsic meaning; these conditions may also restrict $\Lambda\Lambda'$.

Simple structure. These are conditions proposed by Thurstone for choosing a matrix out of the class $\Lambda\theta$ that will have particular psychological meaning. If $\lambda_{ia} = 0$, then the ath factor does not enter into the ith test. The general idea of "simple structure" is that many tests should not depend on all the factors when the factors have real psychological meaning. This suggests that given a Λ one should consider all rotations, that is, all matrices $\Lambda\theta$, where θ is orthogonal, and choose the one giving most 0 coefficients. This matrix can be considered as giving the simplest structure and presumably the one with most meaningful psychological interpretation. It should be remembered that the psychologist can construct his tests so that they depend on the factors in different ways.

If we do not require $\mathcal{E}ff' = I$, then

$$\mathcal{E}(X - \mu)(X - \mu)' = \Sigma + \Lambda M\Lambda', \tag{5.20}$$

where $M = \mathcal{E}ff'$. Then $\Lambda^* = \Lambda A$ and $M^* = A^{-1}M(A^{-1})'$ also satisfies (5.20). The indeterminacy here is indicated by the nonsingular matrix A. Thurstone has suggested simple structure as a means of identification in this case also. Of course, one needs to add a normalization on each component of f or on each column of Λ (as well as an ordering of the columns of Λ).

Thurstone (p. 335 of [24]) suggests that the matrix Λ should be chosen so that there is a submatrix of Λ (obtained by deleting rows of Λ) say $\bar{\Lambda}$ with the following properties: (1) Each row of $\bar{\Lambda}$ should have at least one zero element. (2) Each column of $\bar{\Lambda}$ should have zero elements in at least m rows and these rows should be linearly independent. (It should be pointed out that the desired linear independence is impossible because these rows have zero elements in a given column out of m columns and hence the submatrix of these rows can have maximum rank of $m - 1$.) (3) For every pair of columns of $\bar{\Lambda}$ there should be several rows in which one coefficient is zero and one is nonzero. (4) For every pair of columns in $\bar{\Lambda}$ a large proportion of rows should have two zero coefficients (if $m \geqq 4$). (5) For every pair of columns of $\bar{\Lambda}$ there should preferably be only a small number of rows with two nonzero coefficients.

It is extremely difficult to study the adequacy of these conditions to affect identification. Reiersøl [19] has investigated these conditions, modified a bit. He assumes that there are at least m zero elements in each column of Λ. Let $\Lambda^{(a)}$ $(a = 1, \cdots, m)$ be the submatrix of Λ that has zero elements in the ath column. Reiersøl further assumes that

(i) the rank of $\Lambda^{(a)}$ is $m - 1$, (ii) the rank of each submatrix obtained by deleting a row of $\Lambda^{(a)}$ is $m - 1$, and (iii) the addition to $\Lambda^{(a)}$ of any row of Λ not contained in $\Lambda^{(a)}$ increases the rank to m. Then if $\Lambda\Lambda'$ is identified, a necessary and sufficient condition for the identification of Λ is that Λ does not contain any other submatrices satisfying (i), (ii), and (iii).

Zero elements in specified positions. Here we consider a set of conditions that requires of the investigator more *a priori* information. He must know that some tests do not depend on some factors. In this case the conditions are that $\lambda_{ia} = 0$ for certain pairs (i, a); that is, that the ath factor does not affect the ith test score. In this case we do not assume that $\mathcal{E}ff' = I$. These conditions are similar to some used in econometric models. The coefficients of the ath column are identified except for multiplication by a scale factor if (A) there are at least $m - 1$ zero elements and (B) the rank of $\Lambda^{(a)}$ is $m - 1$ (see [13]).

It will be seen that there are m normalizations and a minimum of $m(m - 1)$ zero conditions. This is equal to the number of elements of A. If there are more than $m - 1$ zero elements specified in one or more columns of Λ, then there may be more conditions than are required to take out the indeterminacy in ΛA; in this case the conditions may restrict $\Lambda M\Lambda'$.

Local identification. We can ask the question, when we suppose there is a Σ and a Λ satisfying $\Psi = \Sigma + \Lambda\Lambda'$ and some other conditions such as $\Lambda'\Sigma^{-1}\Lambda$ being diagonal, is there another pair of such matrices in the neighborhood of Σ, Λ? In other words, if we change Σ and Λ by small amounts, does $\Sigma + \Lambda\Lambda'$ necessarily change? If $\Sigma + \Lambda\Lambda'$ does change, then we say that Σ and Λ are locally identified. We can give a sufficient condition for this.

THEOREM 5.9. *Let* $\Phi = \Sigma - \Lambda(\Lambda'\Sigma^{-1}\Lambda)^{-1}\Lambda'$. *If* $|\phi_{ij}^2| \neq 0$, *then* Σ *and* Λ *are locally identified under the restriction that* $\Lambda'\Sigma^{-1}\Lambda$ *is diagonal and the nondiagonal elements are different and arranged in descending order of size.*

PROOF. Let $\Psi = \Sigma + \Lambda\Lambda'$. Then any pair of matrices Σ^*, Λ^* satisfying $\Psi = \Sigma^* + \Lambda^*\Lambda^{*\prime}$ and $\Lambda^{*\prime}\Sigma^{*-1}\Lambda^*$ diagonal must also satisfy

$$\Lambda^*(I + \Gamma^*) = \Psi\Sigma^{*-1}\Lambda^*, \tag{5.21}$$

$$\operatorname{diag}\Sigma^* = \operatorname{diag}(\Psi - \Lambda^*\Lambda^{*\prime}), \tag{5.22}$$

$$\Lambda^{*\prime}\Sigma^{*-1}\Lambda^* = \Gamma^* \tag{5.23}$$

and the condition that Γ^* is diagonal. As will be seen later, the above equations are analogous to a set of equations defining some estimates. These equations define Λ^* and Σ^* implicitly. We shall show that from these equations one can find the set of partial derivatives $(\partial\sigma_{ii}^*)/(\partial\psi_{jk})$, $(\partial\lambda_{ia}^*)/(\partial\psi_{jk})$. Under the conditions of the theorem the matrix of partial derivatives is of maximum rank (equal to the number of elements in Σ, Λ); this is proved in section 12. The Taylor's series expansion for Σ^* and Λ^* in terms of Ψ is

$$(\Sigma^* - \Sigma, \Lambda^* - \Lambda) = L(\Psi^* - \Psi) \tag{5.24}$$

where L is a linear function. The right-hand side is zero if and only if the left-hand side is zero. Q.E.D.

In a sense the study of identifiability is of more relevance than the study of solvability, for identification requires that the investigator specify some features of the model and he wants to know how to do this. As far as solvability goes, in principle, he either has it or he does not, and there is nothing for him to do about it.

6. Problems of the population: Determination of the structure (III)

The study of solvability and identification implies methods of solving for the structure, given the population of the observables. If the conditions of theorem 5.1 are satisfied, then the communalities can be determined as indicated in the proof of that theorem; this determines $\Lambda\Lambda' = C$, say. Let $\Lambda = (\lambda^{(1)}\lambda^{(2)} \cdots \lambda^{(m)})$, where $\lambda^{(\alpha)}$ is the αth column of Λ. Then

$$C = \lambda^{(1)}\lambda^{(1)\prime} + \lambda^{(2)}\lambda^{(2)\prime} + \cdots + \lambda^{(m)}\lambda^{(m)\prime}. \tag{6.1}$$

In many cases one determines the $\lambda^{(\alpha)}$'s successively. After $\lambda^{(1)}$ is found, we define $C^{(1)} = C - \lambda^{(1)}\lambda^{(1)\prime} = \lambda^{(2)}\lambda^{(2)\prime} + \cdots + \lambda^{(m)}\lambda^{(m)\prime}$, and proceed to find $\lambda^{(2)}$. In turn we define $C^{(\alpha)} = C^{(\alpha-1)} - \lambda^{(\alpha)}\lambda^{(\alpha)\prime}$ and find $\lambda^{(\alpha+1)}$. The methods depend on the identification conditions.

(*a*) *Triangularity conditions.* Since the first components of $\lambda^{(2)}, \cdots, \lambda^{(m)}$ are zero, the first column of C is $\lambda_{11}\lambda^{(1)}$; λ_{11} is determined from $c_{11} = \lambda_{11}^2$ and the rest of $\lambda^{(1)}$ is found from the first column of C. The matrix $C^{(1)} = C - \lambda^{(1)}\lambda^{(1)\prime}$ has only 0's in the first row and column; since the first two components of $\lambda^{(3)}, \cdots, \lambda^{(m)}$ are zero, the second column of $C^{(1)}$ is $\lambda_{22}\lambda^{(2)}$; this determines $\lambda^{(2)}$. In turn $\lambda^{(3)}, \cdots, \lambda^{(m)}$ are found similarly.

(*b*) *General triangularity conditions.* Let

$$F = B'\Lambda = \begin{pmatrix} f_{11} & 0 & \cdots & 0 \\ f_{21} & f_{22} & \cdots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ f_{m1} & f_{m2} & \cdots & f_{mm} \end{pmatrix} = (f^{(1)} \cdots f^{(m)}), \tag{6.2}$$

$$B = (b^{(1)} \cdots b^{(m)}). \tag{6.3}$$

Then $CB = \Lambda F' = \lambda^{(1)}f^{(1)\prime} + \cdots + \lambda^{(m)}f^{(m)\prime}$ and $B'CB = FF' = f^{(1)}f^{(1)\prime} + \cdots + f^{(m)}f^{(m)\prime}$. These two matrix equations can be written

$$Cb^{(\alpha)} = \lambda^{(1)}f_{\alpha 1} + \cdots + \lambda^{(\alpha)}f_{\alpha\alpha}, \qquad \alpha = 1, \cdots, m, \tag{6.4}$$

$$b^{(\beta)\prime}Cb^{(\alpha)} = f_{\beta 1}f_{\alpha 1} + \cdots + f_{\beta\beta}f_{\alpha\beta}, \qquad \beta \leqq \alpha = 1, \cdots, m. \tag{6.5}$$

The first column of CB is $Cb^{(1)} = \lambda^{(1)}f_{11}$, and the first element of $B'CB$ is $b^{(1)\prime}Cb^{(1)} = f_{11}^2$; we determine f_{11} and $\lambda^{(1)}$ from these, which only involve $b^{(1)}$. The second column of CB is $Cb^{(2)} = \lambda^{(1)}f_{21} + \lambda^{(2)}f_{22}$ and two more elements of $B'CB$ are $b^{(1)\prime}Cb^{(2)} = f_{11}f_{21}$ and $b^{(2)\prime}Cb^{(2)} = f_{21}^2 + f_{22}^2$; we find f_{21}, f_{22}, and $\lambda^{(2)}$ from these which involve only the first two columns of B. In turn we find each column of Λ; the αth column only requires use of the first α columns of B.

There is an alternative method for finding $\lambda^{(2)}$ after $\lambda^{(1)}$ is found. Let $C^{(1)} = C - \lambda^{(1)}\lambda^{(1)\prime} = \lambda^{(2)}\lambda^{(2)\prime} + \cdots + \lambda^{(m)}\lambda^{(m)\prime}$; then $C^{(1)}b^{(2)} = \lambda^{(2)}f_{22}$ and $b^{(2)\prime}C^{(1)}b^{(2)} = f_{22}^2$. In turn we define $C^{(\alpha)} = C^{(\alpha-1)} - \lambda^{(\alpha)}\lambda^{(\alpha)\prime}$ and find $\lambda^{(\alpha+1)}$.

(*c*) *Diagonality of* $\Lambda'\Lambda$. Let $d_1, \cdots, d_m$ be the nonzero roots of $|C - dI| = 0$, ordered in descending order, and let $\bar{\lambda}^{(\alpha)}$ be the corresponding vectors satisfying

$(C - d_\alpha I)\bar{\lambda}^{(\alpha)} = 0$ and $\bar{\lambda}^{(\alpha)\prime}\bar{\lambda}^{(\alpha)} = d_\alpha$. (It follows that $\bar{\lambda}^{(\alpha)\prime}\bar{\lambda}^{(\beta)} = 0$, $\alpha \neq \beta$.) If $\bar{\Lambda} = (\bar{\lambda}^{(1)} \cdots \bar{\lambda}^{(m)})$ and $D = (d_\alpha \delta_{\alpha\beta})$, then the equations can be written

$$C\bar{\Lambda} = \bar{\Lambda}D, \tag{6.6}$$

$$\bar{\Lambda}'\bar{\Lambda} = D. \tag{6.7}$$

These equations (and the fact that D is diagonal with ordered elements) determine $\bar{\Lambda}$ and D uniquely. Since $\bar{\Lambda} = \Lambda$ satisfies the equations ($C\Lambda = \Lambda\Lambda'\Lambda = \Lambda(\Lambda'\Lambda)$), it is the unique solution.

(*d*) *Diagonality of* $\Lambda'\Sigma^{-1}\Lambda$. Let $d_1, \cdots, d_m$ be the nonzero roots of $|C\Sigma^{-1} - dI| = 0$ (that is, of $|C - d\Sigma| = 0$) ordered in descending order, and let $\bar{\lambda}^{(\alpha)}$ be the corresponding vectors satisfying $(C\Sigma^{-1} - d_\alpha I)\bar{\lambda}^{(\alpha)} = 0$ and $\bar{\lambda}^{(\alpha)\prime}\Sigma^{-1}\bar{\lambda}^{(\alpha)} = d_\alpha$. These equations can be summarized as

$$C\Sigma^{-1}\bar{\Lambda} = \bar{\Lambda}D, \tag{6.8}$$

$$\bar{\Lambda}'\Sigma^{-1}\bar{\Lambda} = D. \tag{6.9}$$

These equations (for $\bar{\Lambda}$ and D) have the unique solution $\bar{\Lambda} = \Lambda$ and $D = \Lambda'\Sigma^{-1}\Lambda$.

It will be seen later that there is a relation between a method of estimation and a method of determining the structure from the population. However, several methods of estimation can be derived without the motivation of finding an analogue to a method for the population.

7. Problems of statistical inference: Methods of estimation (IV)

7.1. *Preliminary remarks.* We now consider drawing a sample of N observations on X, where $X = \Lambda f + U + \mu$, where f has the distribution $N(0, I)$ and U has the distribution $N(0, \Sigma)$; that is, N observations from $N(\mu, \Sigma + \Lambda\Lambda')$. Let the observations be $x_1, \cdots, x_n$. In all methods of estimation μ is estimated by

$$\bar{x} = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha, \tag{7.1}$$

which is the maximum likelihood estimate of μ. The estimation of Σ and Λ is based upon

$$A = \frac{1}{N}\sum (x_\alpha - \bar{x})(x_\alpha - \bar{x})' = \frac{1}{N}\Big[\sum x_\alpha x_\alpha' - N\bar{x}\bar{x}'\Big]. \tag{7.2}$$

As is well known, $\frac{N}{N-1}A$ is an unbiased estimate of the covariance matrix of X.

We shall now consider a number of estimation methods for Λ and Σ. Later we shall consider estimation methods when f is not considered random, but f_α is a vector of parameters for the αth individual.

7.2. *Maximum likelihood estimates for random factor scores when* $\Lambda\Lambda'$ *is unrestricted.* Maximum likelihood estimates were derived by Lawley [14] for the case of random factor scores when the restriction on the parameters is that $\Lambda'\Sigma^{-1}\Lambda$ is diagonal (and the diagonal elements are ordered in descending order of size). As was seen earlier this re-

striction merely takes out the indeterminacy of the rotation in Λ. The logarithm of the likelihood function for the sample is

$$(7.3)\quad -\tfrac{1}{2}pN\log(2\pi) - \tfrac{1}{2}N\log|\Sigma^*+\Lambda^*\Lambda^{*\prime}| - \tfrac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha-\mu^*)'(\Sigma^*+\Lambda^*\Lambda^{*\prime})^{-1}(x_\alpha-\mu^*)$$
$$= -\tfrac{1}{2}pN\log(2\pi) - \tfrac{1}{2}N\log|\Sigma^*+\Lambda^*\Lambda^{*\prime}| - \tfrac{1}{2}\operatorname{tr}[NA(\Sigma^*+\Lambda^*\Lambda^{*\prime})^{-1}] - \tfrac{1}{2}N(\bar{x}-\mu^*)'(\Sigma^*+\Lambda^*\Lambda^{*\prime})^{-1}(\bar{x}-\mu^*),$$

where we write μ^*, Σ^*, and Λ^* to denote that these are mathematical variables. It will be noticed that replacing Λ^* by $\Lambda^*\theta$, where θ is orthogonal, does not change the likelihood function. Thus if we find μ^*, Σ^*, and Λ^* to maximize the likelihood functon, then μ^*, Σ^* and $\Lambda^*\theta$ will also maximize it. The restriction that $\Lambda^{*\prime}\Sigma^{*-1}\Lambda^*$ be diagonal is a convenience here to make the maximizing variables unique (for almost all samples).

When μ^* is set equal to $\bar{x}$, the last term on the right of (7.3) vanishes. It is easy to verify that (for almost all samples) the likelihood function is maximized when the derivatives (subject to $\Lambda^{*\prime}\Sigma^{*-1}\Lambda^*$ being diagonal) are set equal to zero. The resulting equations (after considerable algebraic manipulation) are

$$(7.4)\quad \hat{\Lambda}(1+\hat{\Gamma}) = A\hat{\Sigma}^{-1}\hat{\Lambda},$$

$$(7.5)\quad \operatorname{diag}\hat{\Sigma} = \operatorname{diag}(A-\hat{\Lambda}\hat{\Lambda}'),$$

$$(7.6)\quad \hat{\Gamma} = \hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda},$$

$$(7.7)\quad \operatorname{nondiag}\hat{\Gamma} = \operatorname{nondiag} 0,$$

where diag B indicates the diagonal matrix formed from the diagonal elements of B and nondiag $B = B -$ diag B. Equation (7.4) can also be written

$$(7.8)\quad \hat{\Lambda}\hat{\Gamma} = (A-\hat{\Sigma})\hat{\Sigma}^{-1}\hat{\Lambda}.$$

These equations may be compared with (6.8) and (6.9). It is seen that (7.8), (7.6) and (7.7) are similar to equations defining the characteristic vectors and roots of A in the metric of $\hat{\Sigma}$. $A-\hat{\Sigma}$ is the sample analogue of C. It is assumed that the m largest roots are positive.

The above equations are practically impossible to solve algebraically. Lawley [14] suggests an iterative procedure which involves approximating $\hat{\Sigma}$, then solving for $\hat{\Lambda}$, then using this in (7.5) to get a new approximation for $\hat{\Sigma}$, etc. In this paper we shall not discuss in detail computational procedures for any estimates; we hope to consider these in a later paper.

7.3. *Maximum likelihood estimates for random factor scores when* $\Lambda\Lambda'$ *is unrestricted and* $\Sigma = \sigma^2 I$*; principal components.* The assumption that $\Sigma = \sigma^2 I$, that is, that Σ is a diagonal matrix with all diagonal elements equal is not an assumption that would ordinarily be suitable, but the assumption leads to an estimate of Λ that is closely related to other methods we discuss. Here

$$(7.9)\quad \Gamma = \Lambda'\Sigma^{-1}\Lambda = \frac{1}{\sigma^2}\Lambda'\Lambda.$$

The condition that Γ is diagonal is equivalent to the condition that $\Lambda'\Lambda$ is diagonal. The equations defining the maximum likelihood estimates are

$$\hat{\Lambda}(\hat{\Gamma}+I) = A\left(\frac{1}{\hat{\sigma}^2}I\right)\hat{\Lambda}, \tag{7.10}$$

$$p\hat{\sigma}^2 = \operatorname{tr}(A-\hat{\Lambda}\hat{\Lambda}'), \tag{7.11}$$

$$\hat{\Gamma} = \hat{\Lambda}'\left(\frac{1}{\hat{\sigma}^2}I\right)\hat{\Lambda}, \tag{7.12}$$

$$\text{nondiag } \hat{\Gamma} = \text{nondiag } 0. \tag{7.13}$$

Comparison of these equations with (7.4) to (7.7) shows the effect of assuming $\Sigma = \sigma^2 I$. We can write the above equations by letting $H = \hat{\sigma}^2(\hat{\Gamma} + I)$ as

$$\hat{\Lambda}H = A\hat{\Lambda}, \tag{7.14}$$

$$p\hat{\sigma}^2 = \operatorname{tr}(A - \hat{\Lambda}\hat{\Lambda}'), \tag{7.15}$$

$$H = \hat{\Lambda}'\hat{\Lambda} + \hat{\sigma}^2 I, \tag{7.16}$$

$$\text{nondiag } H = \text{nondiag } 0. \tag{7.17}$$

Since $\operatorname{tr}(A - \hat{\Lambda}\hat{\Lambda}') = \operatorname{tr} A - \operatorname{tr}\hat{\Lambda}\hat{\Lambda}' = \operatorname{tr} A - \operatorname{tr}\hat{\Lambda}'\hat{\Lambda} = \operatorname{tr} A - \operatorname{tr}(H - \hat{\sigma}^2 I) = \operatorname{tr} A - \operatorname{tr} H + m\hat{\sigma}^2$, we have

$$\hat{\sigma}^2 = \frac{1}{p-m}(\operatorname{tr} A - \operatorname{tr} H). \tag{7.18}$$

Now let us see the relation of the above equation to those defining the characteristic roots and vectors of A. Let the solutions to $|A - dI| = 0$ be $d_1 > d_2 > \cdots > d_p$, and let $l_1, \cdots, l_p$ be the corresponding characteristic vectors [that is, solutions to $(A - d_j I)l_j = 0$] normalized by $l_j'l_j = 1$. Let D be the $m \times m$ diagonal matrix with $d_1, \cdots, d_m$ as diagonal elements and let $L = (l_1, \cdots, l_m)$. Then

$$AL = LD, \tag{7.19}$$

$$L'L = I. \tag{7.20}$$

These equations define D and L uniquely (with the condition that the elements of D are the largest possible). Thus $D = H$ and $L\Delta = \hat{\Lambda}$, where Δ is diagonal. Then $(p - m)\hat{\sigma}^2 = \operatorname{tr} A - \operatorname{tr} H = \sum_1^p d_i - \sum_1^m d_i = \sum_{m+1}^p d_i$. Also

$$H - \hat{\sigma}^2 I = \hat{\Lambda}'\hat{\Lambda} = \Delta' L' L \Delta = \Delta^2. \tag{7.21}$$

Thus the αth diagonal element of Δ, say δ_α, is $\sqrt{d_\alpha - \hat{\sigma}^2}$, and $\hat{\lambda}^{(\alpha)} = \sqrt{d_\alpha - \hat{\sigma}^2}\, l_\alpha$. The characteristic vectors l_α are known as the *principal components* of A. We see here that these are proportional to the maximum likelihood estimates of Λ in our model when $\Sigma = \sigma^2 I$. Hotelling [11] suggested this method when $\Sigma = 0$, or rather when Σ is very small; his point of view was that X had an arbitrary normal distribution and Λf should account for most of the variability of X. For our model we should consider his estimate of Λ as $LD^{1/2}$.

7.4. *Thomson's modification of the principal component method for random factor scores*

when $\Lambda\Lambda'$ is unrestricted. For convenience here we require $\Lambda'\Lambda$ to be diagonal. The equations are

$$\hat{\Lambda}J = (A - \hat{\Sigma})\hat{\Lambda}\,, \tag{7.22}$$

$$\operatorname{diag}\hat{\Sigma} = \operatorname{diag}(A - \hat{\Lambda}\hat{\Lambda}')\,, \tag{7.23}$$

$$J = \hat{\Lambda}'\hat{\Lambda}\,, \tag{7.24}$$

$$\operatorname{nondiag} J = \operatorname{nondiag} 0\,. \tag{7.25}$$

Given $\hat{\Sigma}$, the characteristic vectors of $A - \hat{\Sigma}$ corresponding to the largest characteristic roots constitute the columns of $\hat{\Lambda}$ (normalized according to the diagonal elements of (7.24), that is, the corresponding characteristic roots). Thus the Thomson method [21] is essentially the method of principal components applied to $A - \hat{\Sigma}$.

This method can be compared to the maximum likelihood method by seeing that the maximum likelihood method involves the characteristic vectors and roots of $A - \hat{\Sigma}$ in the metric of $\hat{\Sigma}$.

7.5. *The centroid method.* This method is based on the algebra used to find Λ from C when Λ is restricted by $B'\Lambda$ being triangular (see section 6). Let $\hat{\Sigma}_0$ be an initial approximation to $\hat{\Sigma}$ and let $\hat{C}_0 = A - \hat{\Sigma}_0$. In applying the algebra described in section 6 we choose the columns of B, say B_0, in a way that is apparently suitable for this $\hat{C}_0$. The first row of B_0' is $b_0^{(1)\prime} = (1, 1, \cdots, 1)$; then an element of $\hat{C}_0 b_0^{(1)}$ is the sum of the elements of that row of $\hat{C}_0$ and $b_0^{(1)\prime} C_0 b_0^{(1)}$ is the sum of all elements of $\hat{C}_0$. We form $\hat{C}_0^{(1)} = \hat{C}_0 - \hat{\lambda}_0^{(1)}\hat{\lambda}_0^{(1)\prime}$, and now apply $b_0^{(2)}$. The elements of this vector are 1 or -1. They are chosen so as to make $b_0^{(2)\prime}\hat{C}_0^{(1)}b_0^{(2)}$ as large as possible. The computation of $\hat{C}_0^{(1)}b_0^{(2)}$ is easy because only addition and subtraction of elements of $\hat{C}_0^{(1)}$ are involved. In turn $\hat{C}_0^{(a)} = \hat{C}_0^{(a-1)} - \hat{\lambda}_0^{(a)}\hat{\lambda}_0^{(a)\prime}$ is computed, and then $\hat{\lambda}_0^{(a+1)} (a = 2, \cdots, m - 1)$. Then $\hat{\Lambda}_0 = (\hat{\lambda}_0^{(1)} \cdots \hat{\lambda}_0^{(m)})$ is a first approximation to the estimate of $\hat{\Lambda}$. Next $A - \hat{\Lambda}_0\hat{\Lambda}_0'$ is computed, and the diagonal elements of this matrix (if nonnegative) are taken for $\hat{\Sigma}_1$. Then, the same procedure is followed to obtain $\hat{\Lambda}_1$, another approximation to the estimate of Λ. The matrix taken for B, say B_1, need not be the same as B_0 (except for the first column). In turn $\hat{\Sigma}_i$ and $\hat{\Lambda}_i$ are computed until $A - \hat{\Lambda}_i\hat{\Lambda}_i'$ is a close enough approximation to $\hat{\Sigma}_i$.

In a sense the centroid method is an approximation to Thomson's modification of the principal components method. In that method the first column of $\hat{\Lambda}$ is the characteristic vector of $A - \hat{\Sigma}$ corresponding to the largest characteristic root. This vector is proportional to the normalized vector y (that is, $y'y = 1$) that maximizes $y'(A - \hat{\Sigma})y$, and y satisfies $(A - \hat{\Sigma})y = J_1 y$, where J_1 is the largest characteristic root of $A - \hat{\Sigma}$. If the elements of y are about equal, then y is approximately proportional to $b^{(1)}$, the first column of B, and hence $J_1 y$ is approximately proportional to the first vector of $\hat{\Lambda}$ found by the centroid. Similarly if the second characteristic vector of $A - \hat{\Sigma}$ is approximately proportional to $b^{(2)}$, then it is also approximately proportional to the second column of $\hat{\Lambda}$ by the centroid method. We can say that the centroid method approximates the principal components method by trying to use vectors with elements ± 1 as the characteristic vectors of $A - \hat{\Sigma}$.

The big advantage of the centroid method is the ease of computation. Accordingly, it is the most used method.

7.6. *Maximum likelihood estimates for random factor scores when Λ is identified by specified zero elements.* In this case we have $\mathcal{E}ff' = M$, where M is not required to be

diagonal. However, we require the diagonal elements to be unity. Certain coefficients of Λ are required to be zero, say

$$\lambda_{i\alpha} = 0, i = i(1, \alpha), \cdots, i(p_\alpha, \alpha), \alpha = 1, \cdots, m. \tag{7.26}$$

In the αth column of Λ, there are p_α zero elements and these are in rows numbered $i(1, \alpha), \cdots, i(p_\alpha, \alpha)$. We assume that these conditions effect identification. We can now apply the method of maximum likelihood. We write down the resulting equations, inserting another unknown $p \times m$ matrix J (essentially Lagrange multipliers) which has zero elements where Λ does not; that is,

$$j_{i\alpha} = 0, i \neq i(1, \alpha), \cdots, i(p_\alpha, \alpha), \alpha = 1, \cdots, m. \tag{7.27}$$

The equations are

$$\operatorname{diag} \hat{\Sigma} = \operatorname{diag} (A^{-}\hat{\Lambda}\hat{M}\hat{\Lambda}'), \tag{7.28}$$

$$J'\hat{\Lambda} = 0, \tag{7.29}$$

$$\hat{\Lambda}'\hat{\Sigma}^{-1}A - \hat{\Lambda}' - \hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda}\hat{M}\hat{\Lambda}' = (\hat{M}^{-1} + \hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda})J'\hat{\Sigma}. \tag{7.30}$$

The derivation of these equations is given in section 10. We also consider in more detail a special case when $m = 2$. The above equations cannot be solved algebraically, but iteration methods can be devised.

7.7. *Estimates for nonrandom factor scores when $\Lambda\Lambda'$ is unrestricted.* We now consider $x_\alpha (\alpha = 1, \cdots, N)$ to be an observation on

$$X_\alpha = \Lambda f_\alpha + U + \mu, \tag{7.31}$$

where f_α is a fixed vector. Then the expected value of X_α is

$$\mathcal{E}X_\alpha = \Lambda f_\alpha + \mu, \tag{7.32}$$

and the covariance matrix is

$$\mathcal{E}(X_\alpha - \mathcal{E}X_\alpha)(X_\alpha - \mathcal{E}X_\alpha)' = \Sigma. \tag{7.33}$$

This model is similar to the usual model for least squares (or linear regression) except that here the "independent variates," the f_α, are unknown; the f_α are also parameters.

In one terminology Λ, μ and Σ are considered "structural parameters" because they affect all the random variables, and the f_α are considered "incidental parameters" because each f_α affects only one X_α. The problem of estimating Λ is essentially equivalent to estimating linear equations on the "systematic parts" of X_α. Let $\mathcal{E}X_\alpha = \xi_\alpha$. The hypothesis that ξ_α is of the form $\xi_\alpha = \Lambda f_\alpha + \mu$ is equivalent to the hypothesis that $P\xi_\alpha = \gamma$ where P is a $(p - m) \times p$ matrix such that $P\Lambda = 0$ and $P\mu = \gamma$ (that is, that ξ_α satisfies $p - m$ linear equations).

If we assume U has a normal distribution, the likelihood function is

$$\frac{1}{(2\pi)^{pN/2}|\Sigma|^{N/2}} \exp\left[-\frac{1}{2}\sum_\alpha (x_\alpha - \Lambda f_\alpha - \mu)'\Sigma^{-1}(x_\alpha - \Lambda f_\alpha - \mu)\right] \tag{7.34}$$

$$= \frac{1}{(2\pi)^{pN/2}} \prod_{i=1}^{k} \frac{1}{\sigma_{ii}^{N/2}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N} \frac{(x_{i\alpha} - \Sigma_\nu \lambda_{i\nu} f_{\nu\alpha} - \mu_i)^2}{\sigma_{ii}}\right].$$

The likelihood function does not have a maximum. To show this, let $\mu_1 = 0$, $\lambda_{11} = 1$, $\lambda_{1\nu} = 0$, $\nu \neq 1$, $f_{1\alpha} = x_{1\alpha}$. Then the first term in the product in (7.34) is $\sigma_{11}^{-N/2}$. As

$\sigma_{11} \to 0$, this term is unbounded. Thus, the likelihood function has no maximum, and maximum likelihood estimates do not exist. It might be observed that in [15] Lawley obtained some estimation equations by setting equal to zero the derivatives of the likelihood function; it is not clear, however, whether these equations define even a relative maximum, and they obviously cannot define an absolute maximum. (Lawley applied an iterative method for these equations to some data and found that a σ_{ii} tended towards zero.)

While we cannot apply the method of maximum likelihood to the distribution of $x_1, \cdots, x_n$ to find estimates of all parameters, we can apply the method to the distribution of $A = \frac{1}{N}\sum_{\alpha} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$ to find estimates of Λ and Σ. The distribution of A is the noncentral Wishart distribution [3] and depends on Σ and

$$(7.35)\quad \frac{1}{N}\sum_{\alpha=1}^{N}\Big(\mathcal{E}x_\alpha - \frac{1}{N}\sum \mathcal{E}x_\beta\Big)\Big(\mathcal{E}x_\alpha - \frac{1}{N}\Sigma\mathcal{E}x_\beta\Big)' = \frac{1}{N}\Lambda\sum_{\alpha=1}^{N} f_\alpha f_\alpha'\Lambda'$$

when $\sum_{\alpha} f_\alpha = 0$. If $M = \frac{1}{N}\sum_{\alpha} f_\alpha f_\alpha'$, then the matrix is $\Lambda M \Lambda'$; if we require $M = I$, then the matrix is $\Lambda\Lambda'$. With some restrictions on Λ to take out the rotation, Σ and Λ are identified.

The application of the method of maximum likelihood to the distribution of A is detailed in section 11. Of the resulting equations, one set of m is extremely complicated and cannot be solved explicitly. The other equations are similar to the equations obtained by applying the method of maximum likelihood to the case of random factors.

The question arises whether the maximum likelihood estimates for the case of random factors are suitable for the case of nonrandom factors. In section 11 we prove that the estimates based on maximizing the noncentral Wishart likelihood function are asymptotically equivalent to the maximum likelihood estimates for random factors in the sense that $\sqrt{N}$ times the difference of the two respective estimates converges stochastically to zero. It would, therefore, appear that for large samples in the case of nonrandom factors one can use the maximum likelihood estimates for random factors.

Another asymptotic result that is proved in Part II is that under certain suitable identification conditions the asymptotic distribution of the maximum likelihood estimate of Λ for random factors is the same whatever the assumption on the factors.

7.8. *Units of measurement.* In the preceding sections we have considered factor analysis methods applied to covariance matrices. In many cases the unit of measurement of each component of x is arbitrary. For instance, in psychological tests, the unit of scoring has no intrinsic meaning. We now consider how changes in the units of measurement affect the analysis.

Changing the units of measurement means multiplying each component of x by a constant; we are interested in cases where not all of these constants are equal. It would be desirable that when a given test score is multiplied by a constant the factor loadings for the test are multiplied by the same constant and the error variance is multiplied by square of the constant. Suppose $Dx = x^*$, where D is a diagonal matrix and not all the diagonal elements are the same. Then $\mathcal{E}x^* = D\mu = \mu^*$, say, and

$$(7.36)\quad \mathcal{E}(x^* - \mu^*)(x^* - \mu^*)' = D\Psi D = D\Lambda(D\Lambda)' + D\Sigma D = \Psi^*,$$

say. Now represent this as

$$\Psi^* = \Lambda^*\Lambda^{*\prime} + \Sigma^* . \tag{7.37}$$

Clearly Σ^* can be taken as $D\Sigma D$ and $\Lambda^*\Lambda^{*\prime}$ can be taken as $D\Lambda(D\Lambda)'$ (and must be taken this way if Σ and $\Lambda\Lambda'$ are identified), but whether Λ^* can be taken as $D\Lambda$ depends on what kind of restrictions are imposed on Λ and Λ^* to make each unique. If Λ (and Λ^*) is required to have an upper triangular matrix of 0's, then so does $D\Lambda$ and $D\Lambda = \Lambda^*$. If $B'\Lambda$ (and $B'\Lambda^*$) is required to have an upper triangle of 0's, then usually $B'D\Lambda$ will not, and $D\Lambda \neq \Lambda^*$. If $\Lambda'\Lambda$ (and $\Lambda^{*\prime}\Lambda^*$) is required to be diagonal then usually $(D\Lambda)'D\Lambda = \Lambda'D^2\Lambda$ will not, and $D\Lambda \neq \Lambda^*$. If $\Lambda'\Sigma^{-1}\Lambda$ (and $\Lambda^{*\prime}\Sigma^{*-1}\Lambda^*$) is required to be diagonal, then $(D\Lambda)'(D\Sigma D)^{-1}D\Lambda = \Lambda'\Sigma^{-1}\Lambda$ and $D\Lambda = \Lambda^*$.

Now let us see how the estimation methods depend on the units of measurement. Let $Dx_\alpha = x_\alpha^*$. Then $NA^* = \sum (x_\alpha^* - \bar{x}^*)(x_\alpha^* - \bar{x}^*)' = NDAD$. The equations for the maximum likelihood estimates of section 7.2 are then

$$\hat{\Lambda}^*(I + \hat{\Gamma}^*) = DAD\hat{\Sigma}^{*-1}\hat{\Lambda}^* , \tag{7.38}$$

$$\operatorname{diag} \hat{\Sigma}^* = \operatorname{diag} (DAD - \hat{\Lambda}^*\hat{\Lambda}^{*\prime}) , \tag{7.39}$$

$$\hat{\Gamma}^* = \hat{\Lambda}^{*\prime}\hat{\Sigma}^{*-1}\hat{\Lambda}^* , \tag{7.40}$$

$$\operatorname{nondiag} \hat{\Gamma}^* = \operatorname{nondiag} 0 . \tag{7.41}$$

Clearly $\hat{\Lambda}^* = D\hat{\Lambda}$, and $\hat{\Sigma}^* = D\hat{\Sigma}D$ is the solution [when $\hat{\Lambda}$ and $\hat{\Sigma}$ is a solution to (7.4) to (7.7)]. Then the results of this method do not essentially depend on the units of measurement.

The second estimation procedure considered assumes $\Sigma = \sigma^2 I$. In the new units $\Sigma^* = D\Sigma D = \sigma^2 D^2$ which is not proportional to I and therefore, if this method is applicable to Ψ, it is not applicable to Ψ^*.

In the third method the transformed equations are

$$\hat{\Lambda}^* J^* = (DAD - \hat{\Sigma}^*)\hat{\Lambda}^* , \tag{7.42}$$

$$\operatorname{diag} \hat{\Sigma}^* = \operatorname{diag} (DAD - \hat{\Lambda}^*\hat{\Lambda}^{*\prime}) , \tag{7.43}$$

$$J^* = \hat{\Lambda}^{*\prime}\hat{\Lambda}^* , \tag{7.44}$$

$$\operatorname{nondiag} J^* = \operatorname{nondiag} 0 . \tag{7.45}$$

We know that because of (7.44) and (7.45) $\hat{\Lambda}^* \neq D\hat{\Lambda}$, but we can ask the question whether $\hat{\Lambda}^* = D\hat{\Lambda}P$, where P is orthogonal; that is, whether $\hat{\Lambda}^*\hat{\Lambda}^{*\prime} = D\hat{\Lambda}\hat{\Lambda}'D$ (whether $\hat{\Lambda}^*$ defines the same factor space as $D\hat{\Lambda}$). If $\hat{\Lambda}^*\hat{\Lambda}^{*\prime} = D\hat{\Lambda}\hat{\Lambda}'D$, then $\hat{\Sigma}^* = D\hat{\Sigma}D$ and (7.42) can be written

$$(D^{-1}\hat{\Lambda}^*)J^* = (A - \hat{\Sigma})D^2(D^{-1}\hat{\Lambda}^*) . \tag{7.46}$$

This indicates that the diagonal elements of J^* are the m largest roots of

$$|(A - \hat{\Sigma}) - JD^{-2}| = 0 , \tag{7.47}$$

and the columns of $D^{-1}\hat{\Lambda}^*$ are the corresponding vectors satisfying

$$[(A - \hat{\Sigma})D^2 - J_\alpha I]\hat{\lambda}^{(\alpha)} = 0 . \tag{7.48}$$

However, the roots of (7.47) will, in general, not be the roots of $A - \hat{\Sigma}$ and the vectors satisfying (7.48) will not span the same linear subspace as the first m characteristic vec-

tors of $A - \hat{\Sigma}$. Thus changing the units of measurement will change the estimated factor space in the Thomson method.

Now let us consider the centroid method. Since we know that if $B'\Lambda$ has an upper triangle of 0's, then $B'D\Lambda$ in general will not, we ask whether the centroid method applied to $A^* = DAD$ will give $\hat{\Lambda}^*\hat{\Lambda}^{*\prime} = D\hat{\Lambda}\hat{\Lambda}'D$; that is, whether $\hat{\Lambda}^* = D\hat{\Lambda}P$, where P is some orthogonal matrix. In the original metric, we have $\hat{C}B = \hat{\Lambda}F'$, where $\hat{C} = A - \hat{\Sigma}$ and diag $\hat{\Sigma}$ = diag $(A - \hat{\Lambda}\hat{\Lambda}')$. If $\hat{\Lambda}^* = D\hat{\Lambda}P$, then diag $\hat{\Sigma}^*$ = diag $(DAD - D\hat{\Lambda}\hat{\Lambda}'D)$ = diag $D\hat{\Sigma}D$ and $\hat{C}^* = D\hat{C}D$. Then $\hat{\Lambda}^*F^{*\prime} = \hat{C}^*B = D\hat{C}DB$. Let $\hat{\Lambda}^* = D\bar{\Lambda}$. Then $\bar{\Lambda}F^{*\prime} = \hat{C}DB$ and we ask whether $\bar{\Lambda} = \hat{\Lambda}P$. This can be true in general only if $DB(F^{*\prime})^{-1} = B(F')^{-1}P$; that is, only if $DBQ = B$ for some nonsingular Q. In general, this is not true (only if the m columns of B lie in an m-dimensional space spanned by some m characteristic vectors of D). However, in the centroid method the choice of B is left to the investigator, subject to the conditions that the first column is composed of 1's and the other columns have 1's and -1's as elements. Thus, the B^* used for A^* would usually not be the B used for A. Then we would need $DB^*Q = B$. While it is hard to describe exactly how B is chosen by the investigator, we can say roughly that the columns of B are selected as characteristic vectors of C, and thus the columns of $\hat{\Lambda}$ are approximately proportional to the first m characteristic vectors of C. But in the latter case we have shown that the transformation of A to DAD does not transform $\hat{\Lambda}\hat{\Lambda}'$ to $D\hat{\Lambda}\hat{\Lambda}'D$; hence, we can conclude that to the extent that the centroid method approximates the principal components method (applied to C), it does not transform properly with changes of scale of measurement.

In the case where Λ is identified by 0 coefficients in specified positions, $D\Lambda$ satisfies the same conditions and hence $\Lambda^* = D\Lambda$. In estimation J^* has 0's specified in the same positions as in J. It is a straightforward matter to show that $\hat{\Lambda}^* = D\hat{\Lambda}$, $\hat{\Sigma}^* = D\hat{\Sigma}D$, $\hat{M}^* = \hat{M}$, and $\hat{J}^* = D^{-1}\hat{J}$ satisfy (7.28), (7.29) and (7.30) when $A^* = DAD$.

In the case of nonrandom factor scores we suggest applying the method of maximum likelihood to the likelihood of A. It can be seen from results in Part II that $D\hat{\Lambda}$ and $D\hat{\Sigma}D$ satisfy the condition for removing the rotation from $\hat{\Lambda}^*$ (and $\hat{\Lambda}$). The value of the likelihood function at A, Σ, Λ is the same as at DAD, $D\Sigma D$, $D\Lambda$; hence, the maximum for $A^* = DAD$ is at $\hat{\Sigma}^* = D\hat{\Sigma}D$, $\hat{\Lambda}^* = D\hat{\Lambda}$ when the maximum for A is at $\hat{\Sigma}$, $\hat{\Lambda}$.

As has been noted above, the estimation of Λ by the centroid or Thomson's principal components method depends essentially on the units of measurement of the test scores, even though these units may have no intrinsic meaning. A practical remedy to this undesirable indeterminacy is to prescribe a "statistical" unit of measurement. It is customary to let the sample determine the unit of measurement by requiring that each test score have sample variance 1. Thus d_{ii} is taken to be $1/\sqrt{a_{ii}}$. The new matrix is $R = (r_{ij})$, where $r_{ij} = a_{ij}/\sqrt{a_{ii}a_{jj}}$ are the sample correlation coefficients. Besides taking out the indeterminacy, this convention has some other advantages. From the practical point of view it is convenient to have the diagonal elements unity and the other numbers between -1 and $+1$; this makes it easier to find rules of thumb and convenient computational procedures.

The centroid method is an approximation to the modified principal components method. If we compare the equations for the latter with those for the maximum likelihood solution, we see that when $\hat{\Sigma}$ is roughly proportional to I, then the principal component estimates are close to the maximum likelihood estimates, which have certain desirable properties (for example, asymptotic efficiency). If the transformation to test

scores with unit sample variance tends to make the error variances ($\hat{\sigma}_{ii}$) approximately equal, then the efficiency of these procedures is presumably improved.[1]

It might be pointed out that the assumption of section 7.3 that $\sigma^2 = I$ is a little less restrictive than it seems. Suppose one knows the error variances except for a constant of proportionality σ^2. Then $\Sigma = \sigma^2 D^{-2}$, say, where D^{-2} is known. Then we can let $Dx_\alpha = x_\alpha^*$ and apply the principal components method to $A^* = DAD$. It might also be noted that Whittle [28] has treated the nonrandom factor case under the assumption that $\Sigma = \sigma^2 I$ and has obtained a solution for Λ in terms of the principal components of A.

7.9. *Invariance of factor loadings under changes of factor score populations.* Now let us consider the model $X = \Lambda f + \mu + U$, where $\mathcal{E}ff' = M$ is not necessarily required to be the identity, and where f and U are considered random. Of the various ways of identifying Λ (where Σ is identified), consider (*a*) $B'\Lambda$ has an upper triangle of 0's and $M = I$, (*b*) $\Lambda'\Lambda$ is diagonal and $M = I$, (*c*) $\Lambda'\Sigma^{-1}\Lambda$ is diagonal and $M = I$, and (*d*) Λ has specified 0's. Only the last does not involve M.

A mathematical factor analysis is supposed to be a representation of some real population of individuals from which we sample randomly. In defining such a representation it is desirable that at least certain parts of the model do not change even though the population is changed. For example, consider a model for certain mental test scores of a certain population, say, boys of age 16 in New York State. Then consider a subpopulation, say, boys of age 16 in eleventh grade in New York State. Can the same model apply to this subpopulation? To put it another way, if one investigator factor-analyzes the first population and another analyzes the second, what results of the analyses might be common to the two studies (see also [18])?

If the definition of the subpopulation is independent of f and U (that is, does not depend on the factor scores and "errors" including specific factors), then the subpopulation is a miniature of the first and any model for the first furnishes the same model for the second. However, in the example above it would seem reasonable that the subpopulation involves a selection based on the factor scores related to the set of tests considered (as well as other factors).

Let us consider what happens in the above model if f is replaced by g, where $\mathcal{E}g = \gamma$ and $\mathcal{E}(g - \gamma)(g - \gamma)' = P$. Then in the subpopulation

$$X^* = \Lambda g + \mu = U = \Lambda(g - \gamma) + (\mu + \Lambda\gamma) + U\,. \tag{7.49}$$

The investigator is going to represent this as

$$X^* = \Lambda^* f^* + \mu^* + U^*\,, \tag{7.50}$$

where $\mathcal{E}U^* = 0$, $\mathcal{E}f^* = 0$, $\mathcal{E}U^*U^{*\prime} = \Sigma^*$, $\mathcal{E}f^*f^{*\prime} = M^*$ and Λ^* and M^* satisfy the identification conditions.

Let $\mu^* = \mu + \Lambda\gamma$, $U^* = U$, $\Sigma^* = \Sigma$, and $f^* = Q(g - \gamma)$ and $\Lambda^* = \Lambda Q^{-1}$ for some nonsingular Q. It is clear that the columns of Λ^* span the same space as the columns of Λ. If (*d*) is used for identification Q must be diagonal, and each column of Λ^* must be proportional to the corresponding column of Λ; also $q_{ii} = 1/\sqrt{p_{ii}}$. If the normalization

[1] Whittle [28] has suggested that if one assumes the variance of the measurement is proportional to the error variance, then it is reasonable to use the correlation matrix. In the case of nonrandom factor scores, he has assumed $\sum_\nu \lambda_{i\nu}^2 = c\sigma_{ii}$, but finds he is led to principal components of R only in the case of $m = 1$.

of a column of Λ (and Λ^*) is done by a rule involving only that column (for example, by means of making a specified element equal to one), then that column of Λ is equal to that column of Λ^*. It can also be shown that if simple structure effects identification of Λ in the original population, it will in the second and will lead to a Λ^* with proportional columns. In the case of identification by methods (*a*), (*b*), or (*c*) Λ^* is not related as simply to Λ. In each case $M^* = QPQ' = I$. In (*a*) Q also satisfies $B'\Lambda^* = B'\Lambda Q^{-1} = F^*$ (with upper triangle of 0's); in (*b*) $\Lambda^{*\prime}\Lambda^* = (Q^{-1})'\Lambda'\Lambda Q^{-1}$ is diagonal; in (*c*) $\Lambda^{*\prime}\Sigma^{-1}\Lambda^* = (Q^{-1})'\Lambda'\Sigma^{-1}\Lambda Q^{-1}$ is diagonal. In each of these cases Λ^* will in general not be a rotation of Λ. Thus, only if identification does not essentially involve M can one hope that the results of a factor analysis for one population will bear a simple relation to the results for another population if the two populations differ with respect to the factors involved.

There seems to have been a considerable discussion by psychologists of the requirement that $M = I$. Some claim that the orthogonality (that is, lack of correlation) of the factor scores is essential if one is to consider the factor scores as more basic than the test scores. However, if the factor scores are orthogonal for some population, in general they will not be orthogonal for another population or for a subpopulation. Hence, this requirement would seem to lead to a less basic definition of factors.

We might also consider the effect of the use of correlations in factor analysis on the comparability of analyses of different populations. In the original population $\Psi = \Lambda M\Lambda' + \Sigma$ and $R = D\Psi D = (D\Lambda)M(D\Lambda)' + D\Sigma D$, where $d_{ii}^2 = \sum \lambda_{i\nu}m_{\nu\mu}\lambda_{i\mu} + \sigma_{ii}$. In the second population $\Psi^* = \Lambda P\Lambda' + \Sigma$ and $R^* = D^*\Psi^*D^* = (D^*\Lambda)P(D^*\Lambda)' + D^*\Sigma D^*$, where $d_{ii}^{*2} = \sum \lambda_{i\nu}p_{\nu\mu}\lambda_{i\mu} + \sigma_{ii}$. Then the relation of the factor loading matrix of R^* to that of R is further complicated by the premultiplication and postmultiplication of a diagonal matrix that depends on the subpopulation (that is, on P). Thus the use of correlations instead of covariances makes the comparison of factor loadings in two populations more difficult.

A question related to the above is what happens to the analysis if tests are added (or deleted). Let

$$X^* = \Lambda^* f + \mu^* + U^*, \tag{7.51}$$

where X^* is a vector of added test scores. Then the entire set consists of the components of X and X^*. We assume $EU^* = 0$, $EUU^{*\prime} = 0$, $EU^*U^{*\prime} = \Sigma^*$, a diagonal matrix. What identification conditions leave Λ unaffected? The conditions in terms of the entire set of tests are $(\Lambda'\Lambda^{*\prime})B = \Lambda'B + \Lambda^{*\prime}B$ is a triangular matrix in (*a*), $(\Lambda'\Lambda^{*\prime})(\Lambda'\Lambda^{*\prime})' = \Lambda'\Lambda + \Lambda^{*\prime}\Lambda^*$ is diagonal in (*b*),

$$\begin{pmatrix}\Lambda\\ \Lambda^*\end{pmatrix}'\begin{pmatrix}\Sigma & 0\\ 0 & \Sigma^*\end{pmatrix}^{-1}\begin{pmatrix}\Lambda\\ \Lambda^*\end{pmatrix} = \Lambda'\Sigma^{-1}\Lambda + \Lambda^*\Sigma^{*-1}\Lambda^* \tag{7.52}$$

is diagonal in (*c*). In general, these will not be satisfied. In (*d*), however, the restrictions are the same. Thus in the first three cases addition of new tests will lead usually to a rotation of Λ.

7.10. *Asymptotic distributions of estimates.* For any of the estimation procedures described in this paper, it would be desirable to have the distribution of the estimates. However, in general the exact distribution of any set of estimates is virtually impossible to obtain. The best we can hope for is an asymptotic distribution theory for a set of estimates.

In section 12 we prove that the maximum likelihood estimates given in section 7.2 are asymptotically normally distributed if the matrix (ϕ_{ij}^2) is nonsingular, where $\Phi = \Sigma - \Lambda(\Lambda'\Sigma^{-1}\Lambda)^{-1}\Lambda'$; the condition is implied by the condition that Σ and Λ are identified. Some of the asymptotic variances and covariances are given in section 12. Unfortunately, the variances and covariances of the elements of $\hat{\Lambda}$ are so complicated that they cannot be used for all the usual purposes.

The asymptotic variances and covariances of the estimates in section 7.3 have been given by Lawley [16] and the asymptotic normality has also been proved [4]. However, the assumptions underlying this theory ($\Sigma = \sigma^2 I$) are so restrictive that it would appear that the theory is not of much applicability.

In section 12 it is stated that the modified principal component estimates (of section 7.4) are asymptotically normally distributed if (θ_{ij}^2) is nonsingular, where $\Theta = I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda'$. The asymptotic variances and covariances can be found in a fashion similar to those of the maximum likelihood estimates. Again they are very complicated.

The centroid estimates are not defined explicitly in terms of mathematical operations because the investigator chooses B somewhat subjectively. Hence, we cannot define any asymptotic distribution theory. It is possible to formalize the procedure by assuming that $p = k2^m$, where k is an integer, and defining $b^{(1)\prime} = (1, 1, \cdots, 1)$, $b^{(2)}$ consists of $p/2$ 1's and $p/2$ -1's such that $b^{(2)\prime}C^{(1)}b^{(2)}$ is a maximum, $b^{(3)}$ consists of $p/4$ 1's where $b^{(2)}$ has 1's, $p/4$ -1's where $b^{(2)}$ has 1's, etc., such that $b^{(3)\prime}C^{(2)}b^{(3)}$ is a maximum, etc. This procedure, however, is very difficult to study.

In the case of maximum likelihood estimates when Λ is identified by zero elements, the estimates are asymptotically normally distributed. The proof of this theorem (theorem 12.3) is not given because it is extremely complicated.

When the factor scores are nonrandom (section 7.7), it is stated in section 11 that the estimates based on the distribution of A are asymptotically equivalent to the estimates given in section 7.2 in the sense that $\sqrt{N}$ times the difference of two estimates converges asymptotically to zero. Thus, as far as asymptotic normality and asymptotic variances and covariances are concerned, the two methods are equivalent. It follows from theorem 12.1 that these estimates are asymptotically normally distributed.

8. Problems of statistical inference: Tests of hypotheses and determination of the number of factors

8.1. *Test of the hypothesis that the model fits* (V). In the discussion of estimation we have assumed that the model is proper for the relevant data; in particular, we have assumed that m, the number of factors, is known. In this section we consider testing the hypothesis that Ψ, the population covariance matrix, can be written as $\Sigma + \Lambda\Lambda'$, where Λ has a specified number of columns. There are other assumptions of the model, such as normality, linearity of effects of factors, etc., which may be questioned, but we will not consider them here.

One method of obtaining a test of the hypothesis $\Psi = \Sigma + \Lambda\Lambda'$ is to derive the likelihood ratio criterion under the conditions of section 7.2. The likelihood function can be written

$$(8.1)\quad L(A, \Psi, \mu) = (2\pi)^{-pN/2}|\Psi|^{-N/2}\exp\{-[N \operatorname{tr} A\Psi^{-1} + N(\bar{x} - \mu)'\Psi^{-1}(\bar{x} - \mu)]/2\}\ .$$

The alternatives to the hypothesis are that Ψ is any positive definite matrix. Under the

alternative hypotheses $\hat{\mu} = \bar{x}$ and $\hat{\Psi} = A$. Under the null hypothesis $\hat{\mu} = \bar{x}$ and $\hat{\Psi} = \hat{\Sigma} + \hat{\Lambda}\hat{\Lambda}'$, where $\hat{\Sigma}$ and $\hat{\Lambda}$ are defined in section 7.2. The likelihood ratio criterion is

$$\frac{L(A, \hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}', \hat{\mu})}{L(A, A, \hat{\mu})} = \frac{|A|^{N/2} e^{Np/2}}{|\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}'|^{N/2} e^{N \operatorname{tr} A(\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}')^{-1}/2}}. \tag{8.2}$$

The exponent in (8.2) involves

$$\begin{aligned}\operatorname{tr} A(\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}')^{-1} - p &= \operatorname{tr}(A-\hat{\Sigma}-\hat{\Lambda}\hat{\Lambda}')(\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}')^{-1} \\ &= \operatorname{tr}(A-\hat{\Sigma}-\hat{\Lambda}\hat{\Lambda}')[\hat{\Sigma}^{-1} - \hat{\Sigma}^{-1}\hat{\Lambda}(I+\hat{\Gamma})^{-1}\hat{\Lambda}'\hat{\Sigma}^{-1}] \\ &= \operatorname{tr}(A-\hat{\Sigma}-\hat{\Lambda}\hat{\Lambda}')\hat{\Sigma}^{-1} - \operatorname{tr}(A-\hat{\Sigma}-\hat{\Lambda}\hat{\Lambda}') \\ &\qquad \times \hat{\Sigma}^{-1}\hat{\Lambda}(I+\hat{\Gamma})^{-1}\hat{\Lambda}'\hat{\Sigma}^{-1} \\ &= 0\end{aligned} \tag{8.3}$$

because (7.5) implies the diagonal elements of $(A - \hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}')\hat{\Sigma}^{-1}$ are 0 and (7.8) implies $(A - \hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}')\hat{\Sigma}^{-1}\hat{\Lambda} = 0$. Because $|I + PQ| = |I + QP|$,

$$\begin{aligned}|\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}'| &= |\hat{\Sigma}| \cdot |I+\hat{\Lambda}\hat{\Lambda}'\hat{\Sigma}^{-1}| = |\hat{\Sigma}| \cdot |I+\hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda}| \\ &= |\hat{\Sigma}| \cdot |I+\hat{\Gamma}|.\end{aligned} \tag{8.4}$$

It is convenient to consider (-2) times the logarithm of the likelihood ratio criterion which is

$$U_m = N[\log|\hat{\Sigma}| + \log|I+\hat{\Gamma}| - \log|A|]. \tag{8.5}$$

The test procedure is to reject the hypothesis if U_m exceeds a number; this number is chosen to give the desired significance level. While the exact distribution of U_m is not known, the usual asymptotic theory tells us that if (ϕ_{ij}^2) is nonsingular U_m is asymptotically distributed as χ^2 with number of degrees of freedom equal to $C = p(p+1)/2 + m(m-1)/2 - p - pm$.

The diagonal elements of $\hat{\Gamma}$ in section 7.2 are the m largest roots of

$$|A - \hat{\Sigma} - \gamma\hat{\Sigma}| = 0. \tag{8.6}$$

Let $\hat{\gamma}_{m+1}, \cdots, \hat{\gamma}_p$ be the other roots of (8.6). Then it can be shown that

$$\operatorname{tr} A(\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}')^{-1} - p = \sum_{i=m+1}^{p} \hat{\gamma}_i, \tag{8.7}$$

$$|A(\hat{\Sigma}+\hat{\Lambda}\hat{\Lambda}')^{-1}| = \prod_{i=m+1}^{p}(1+\hat{\gamma}_i). \tag{8.8}$$

Thus the criterion is

$$U_m = -N\sum_{i=m+1}^{p}\log(1+\hat{\gamma}_i). \tag{8.9}$$

We can give an intuitive interpretation of this test. $\hat{\Sigma}$ and $\hat{\Lambda}$ are found so that $A - (\hat{\Sigma} + \hat{\Lambda}\hat{\Lambda}')$ is small in a statistical sense, or equivalently so $A - \hat{\Sigma}$ is approximately of rank m. If $A - \hat{\Sigma}$ is approximately of rank m, the smallest $p - m$ roots of (4.6) should be near zero. The criterion measures in a certain way the deviation of the smallest roots from zero. The criterion is approximately $N\sum_{i=m+1}^{p}\frac{\hat{\gamma}_i^2}{2}$.

This test was proposed by Lawley [14]. Bartlett [6] has suggested replacing the factor N in the criterion by $N - (2p + 11)/6 - 2m/3$. It might be noted that to justify the statement that U_m has an asymptotic χ^2-distribution it is necessary to verify that $\hat{\Lambda}$ and $\hat{\Sigma}$ have an asymptotic normal distribution (or an equivalent statement).

We can also use the likelihood ratio criterion if the identification is by elements of Λ specified zero. Then

$$(8.10)\qquad \begin{aligned} |\hat{\Psi}| &= |\hat{\Sigma} + \hat{\Lambda}\hat{M}\hat{\Lambda}'| = |\hat{\Sigma}|\cdot|I + \hat{\Lambda}M\hat{\Lambda}'\hat{\Sigma}^{-1}| \\ &= |\hat{\Sigma}|\cdot|I + \hat{M}\hat{\Gamma}| = |\hat{\Sigma}|\cdot|\hat{\Gamma}^{-1}\hat{K}| = |\hat{\Sigma}|\cdot|\hat{\Gamma}^{-1}|\cdot|\hat{K}|\,, \end{aligned}$$

and by (10.18)

$$(8.11)\qquad \begin{aligned} \operatorname{tr}\hat{\Psi}^{-1}A - p &= \operatorname{tr}(\hat{\Psi}^{-1}A - I) \\ &= \operatorname{tr}[\hat{\Sigma}^{-1}(A - \hat{\Psi}) - \hat{\Sigma}^{-1}\hat{\Lambda}J'\hat{\Psi}] \\ &= \operatorname{tr}\hat{\Sigma}^{-1}(A - \hat{\Psi}) - \operatorname{tr}\hat{\Sigma}^{-1}\hat{\Lambda}J'\hat{\Sigma} - \operatorname{tr}\hat{\Sigma}\hat{\Lambda}J'\hat{\Lambda}M\hat{\Lambda}' \\ &= 0 \end{aligned}$$

because $\operatorname{diag}(A - \hat{\Psi}) = 0$, $J'\hat{\Lambda} = 0$, and $\operatorname{diag}\hat{\Lambda}J' = 0$. Thus in this case

$$(8.12)\qquad U_m = N[\log|\hat{\Sigma}| + \log|\hat{K}| - \log|\hat{\Gamma}| - \log|A|]\,.$$

When the null hypothesis is true, U_m is distributed as χ^2 with number of degrees of freedom equal to $p(p+1)/2 - p - pm - m(m-1)/2$ plus the number of 0's specified in Λ (at least $m(m-1)$).

Bartlett [6], [8] has proposed another test procedure. Let $z_1 > z_2 > \cdots > z_p$ be the roots of

$$(8.13)\qquad |R - zI| = 0\,.$$

The criterion suggested is

$$(8.14)\qquad \left(N - \frac{2p+11}{6} - \frac{2m}{3}\right)\left[(p-m)\log\frac{\sum_{i=m+1}^{p} z_i}{p-m} - \sum_{i=m+1}^{p}\log z_i\right].$$

This criterion is small if the last $p - m$ roots of R are nearly equal. It is difficult to relate this test to the factor analysis model. The test can be expected to be consistent if the population correlation matrix is of the form $\sigma^2 I + \Lambda\Lambda'$; intuitively the test judges whether $R - \hat{\Lambda}\hat{\Lambda}'$ is approximately proportional to I (see [4]). Another difficulty with this procedure is that even its asymptotic distribution is unknown.

There are other ways of deciding whether $A - \hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}'$ is sufficiently small to accept the hypothesis that $\Psi - \Sigma - \Lambda\Lambda'$ is zero. As was noted earlier, it is common practice to apply the centroid method to the correlation matrix. If $\hat{\Sigma}$ and $\hat{\Lambda}$ are now the estimates based on this method, one wants to decide whether the elements of $R - \hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}'$ are sufficiently near zero. Frequently, rules of thumb are used, such as deciding the elements are sufficiently near zero if each element is within .05 of being zero. It is obviously difficult to investigate the theory of such rules.

8.2. *Determination of the number of factors (VI).* In many cases the investigator does not know the number of factors. He may not even be in the position of postulating a specific value of m. In such situations the statistical problem is not one of testing the hypothesis that m is a given number, but rather of determining an appropriate number

of factors. The investigator wants to decide the smallest number of factors such that the corresponding model fits the data.

What is desired is a multiple decision procedure. Such a procedure could be described by a function $h(A)$, that takes on the values $0, 1, \cdots, M$, where M is the maximum number of factors that could be needed. One would like such a function that had some desirable properties. Unfortunately, there is no such procedure for which we have an adequate statistical theory, even an asymptotic theory.

In a situation such as this it is common practice to make a sequence of tests. In this case one might test the hypothesis that $m = m_0$ against the alternative that $m > m_0$; if this hypothesis is rejected, test $m = m_0 + 1$ against the alternative $m > m_0 + 1$, etc. However, even if we know the significance level of each test separately, we do not know the probabilities associated with the entire procedure; that is, if m^* is the true number of factors, we do not know (even asymptotically) what the probability is that our procedure leads to the decision $m = m^*$. Perhaps all that can be said is that the probability of saying $m > m^*$ $(\geqq m_0)$ is not greater than the significance level of the test of $m = m^*$.

Another kind of procedure that might be considered is another sequence of tests, namely test hypotheses $m = m_0$ against the alternative $m = m_0 + 1$, if this is rejected test $m = m_0 + 1$ against $m = m_0 + 2$, etc. In the case of likelihood ratio tests, this would involve the criterion $U_{m_0} - U_{m_0+1}$, then $U_{m_0+1} - U_{m_0+2}$, etc. However, here we do not know the asymptotic distribution of the criterion.

Using a sequence of likelihood ratio tests is computationally difficult. For at each stage one must compute $\hat{\Sigma}$ and $\hat{\Gamma}$ for a value of m and at the next stage one must compute another $\hat{\Sigma}$ and $\hat{\Gamma}$.

In practice *ad hoc* rules are frequently used in dealing with R, such as using a rule of thumb to determine $\hat{\Sigma}$, then using the centroid method to estimate successively columns of Λ until the elements of $R - \hat{\Sigma} - \hat{\Lambda}\hat{\Lambda}'$ are sufficiently small, say within .05 of being zero. We do not know the statistical properties of such procedures. It has also been proposed that after following such an *ad hoc* procedure, the investigator then use the likelihood ratio test to test the hypothesis that m is equal to the number determined by the *ad hoc* procedure. Again we do not have any statistical theory for the procedure.

8.3. *Tests of hypotheses (VII).* There are many hypotheses about Λ and Σ that an investigator might consider. For example, he might be interested in whether a specified λ_{ia} is zero, that is, whether a given factor does not enter a given test. In this connection, he might also want a confidence interval for a specified λ_{ia}. In principle, it is possible to give a large sample procedure in such cases if one has a consistent estimate which is asymptotically normally distributed, if one knows the theoretical asymptotic variance (in terms of Λ and Σ), and if one has consistent estimates of the parameters involved in the asymptotic variance. Unfortunately, in practice this is extremely difficult because the asymptotic variances are complicated functions of the parameters. It might be noted that when Λ is identified by arbitrary mathematical conditions (such as diagonality of $\Lambda'\Sigma^{-1}\Lambda$), these hypotheses do not have much significance.

Another hypothesis that might be of interest is the hypothesis that all factor loadings for a specified test are zero. In our model this is equivalent to the hypothesis that the specified test score is independent of the other test scores. Such a hypothesis can be tested by using the multiple correlation coefficient between the specified test and the other tests.

Another hypothesis is whether a given tetrad difference is zero. This is relevant to the

question of whether some four tests meet the conditions for a single factor model. Although a fair amount of work has been done on this problem, we shall not treat it because we are interested here in problems for a general number of factors.

9. Problems of statistical inference: Estimation of factor scores (VIII)

If one postulates a model where the factor scores are not random, then the observed test score vector for the αth person is an observation on $\Lambda f_\alpha + \mu + U$, where f_α is a vector of parameters. Given a sample of test score vectors, one from each of N individuals, one can ask for estimates of the N factor score vectors, $f_1, \cdots, f_N$. If one postulates a model where the factor scores are random, one can consider the conditional distribution of the set of test score vectors given that the factor score vectors are fixed vectors, $f_1, \cdots, f_N$. Then this conditional distribution is the same as the distribution with nonrandom factor score vectors.

As was indicated in section 7.6, we cannot apply the method of maximum likelihood to the problem of simultaneous estimation of Σ, Λ, μ, and $f_1, \cdots, f_N$. A reasonable procedure seems to be to estimate Σ, Λ and μ and then consider estimating $f_1, \cdots, f_N$ assuming that Σ, Λ and μ are known and are equal to the estimates $\hat{\Sigma}$, $\hat{\Lambda}$ and $\hat{\mu}$. Under the assumption of normality, we can consider the likelihood of $x_1, \cdots, x_N$ (given Λ, Σ, and μ) and maximize it with respect to $f_1, \cdots, f_N$. This is equivalent to minimizing

$$\sum_{i=1}^{p} \frac{\left(x_{i\alpha} - \mu_i - \sum_{\nu=1}^{m} \lambda_{i\nu} f_{\nu\alpha}\right)^2}{\sigma_i^2}, \qquad \alpha = 1, \cdots, N. \tag{9.1}$$

This amounts to a weighted least-squares problem in which $x_{i\alpha} - \mu_i$ are the dependent variates, $\lambda_{i\nu}$ are the independent variates and $f_{\nu\alpha}$ are the unknown coefficients [5]. The estimated vector f_α is

$$\hat{f}_\alpha = (\Lambda'\Sigma^{-1}\Lambda)^{-1}\Lambda'\Sigma^{-1}(x_\alpha - \mu). \tag{9.2}$$

This estimate has the usual properties of a least-squares estimate. It is unbiased and each component of the estimate has minimum variance of all linear unbiased estimates. It will be observed that the coefficients of the estimates depend on Λ and Σ; these would not change if a selection on the basis of factor scores was made.

Another approach has been suggested by Thomson [21]. If $\mathcal{E}ff' = I$, the covariance matrix of X and f is

$$\mathcal{E}\begin{pmatrix} X \\ f \end{pmatrix}\begin{pmatrix} X \\ f \end{pmatrix}' = \begin{pmatrix} \Sigma + \Lambda\Lambda' & \Lambda \\ \Lambda' & I \end{pmatrix}. \tag{9.3}$$

Then the regression of f on X is $\Lambda'(\Sigma + \Lambda\Lambda')^{-1}X = (I + \Gamma)^{-1}\Lambda'\Sigma^{-1}X$. The estimate of f_α is

$$\hat{f}_\alpha = (I + \Gamma)^{-1}\Lambda'\Sigma^{-1}(x_\alpha - \bar{x}). \tag{9.4}$$

The νth component of this estimate is $\gamma_{\nu\nu}/(1 + \gamma_{\nu\nu})$ times the νth component of (9.2) when Γ is diagonal.

One might also ask for an estimate such that $\frac{1}{N}\sum \hat{f}_\alpha \hat{f}_\alpha' = I$. Now let us find estimates that have this property by minimizing $\sum_\alpha (x_\alpha - \bar{x} - \Lambda f_\alpha)'\Sigma^{-1}(x_\alpha - \bar{x} - \Lambda f_\alpha)$

under the restrictions that $\sum f_\alpha f_\alpha' = NI$. We minimize

$$(9.5) \qquad \sum_{\alpha=1}^{N} (x_\alpha - \bar{x} - \Lambda f_\alpha)'\Sigma^{-1}(x_\alpha - \bar{x} - \Lambda f_\alpha) + \operatorname{tr} \Theta\left(\sum_{\alpha=1}^{N} f_\alpha f_\alpha' - NI\right),$$

where Θ is a symmetric matrix of Lagrange multipliers. Then

$$(9.6) \qquad \hat{f}_\alpha = (\Lambda'\Sigma^{-1}\Lambda + \Theta)^{-1}\Lambda'\Sigma^{-1}(x_\alpha - \bar{x}),$$

where $\Lambda'\Sigma^{-1}\Lambda + \Theta$ is the symmetric square root of $\Lambda'\Sigma^{-1}A\Sigma^{-1}\Lambda$. If Λ and Σ are the estimates by the method of section 7.2 then $\Lambda'\Sigma^{-1}A\Sigma^{-1}\Lambda = \Gamma(I + \Gamma)$ and (9.6) is

$$(9.7) \qquad \hat{f}_\alpha = [\Gamma(I + \Gamma)]^{-1/2}\Lambda'\Sigma^{-1}(x_\alpha - \bar{x}).$$

If we want $\Sigma f_\alpha f_\alpha' = NM$, where M is specified, then $\Lambda'\Sigma^{-1}\Lambda + \Theta$ must satisfy

$$(9.8) \qquad (\Lambda'\Sigma^{-1}\Lambda + \Theta)M(\Lambda'\Sigma^{-1}\Lambda + \Theta) = \Lambda'\Sigma^{-1}A\Sigma^{-1}\Lambda.$$

We note that if one assumes $\Sigma = \sigma^2 I$ and that the factor vectors are nonrandom, then $\mathcal{E} X_{i\alpha} - \mu_i = \sum_\nu \lambda_{i\nu} f_{\nu\alpha}$ and the variance of $X_{i\alpha}$ is σ^2. Then the role of tests and individuals can be interchanged. Whittle has shown that under suitable identification conditions (including $\sum_\alpha f_{\nu\alpha} f_{\beta\alpha} = 0$ for $\nu \neq \beta$) the estimates of $f_{\nu\alpha}$ involve the principal components of $\sum_i (x_{i\alpha} - \bar{x}_i)(x_{i\beta} - \bar{x}_i)$.

PART II. PROOFS OF SOME NEW RESULTS

10. Maximum likelihood estimates for random factor scores when Λ is identified by specified zero elements

THEOREM 10.1. *Let $x_1, \cdots, x_n$ be N observations from $N(\mu, \Psi)$, where $\Psi = \Lambda M \Lambda' + \Sigma$. Let $m_{ii} = 1$ and $\lambda_{i\alpha} = 0$, $i = i(1, \alpha), \cdots, i(p_\alpha, \alpha)$, $\alpha = 1, \cdots, m$. Let $NA = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$. The maximum likelihood estimate of μ is $\hat{\mu} = \bar{x}$ and the maximum likelihood estimates of Λ, M and Σ are given by*

$$(10.1) \qquad \operatorname{diag}(A - \hat{\Sigma} - \hat{\Lambda}\hat{M}\hat{\Lambda}') = 0,$$

$$(10.2) \qquad J'\hat{\Lambda} = 0,$$

$$(10.3) \qquad \hat{\Lambda}'\hat{\Sigma}^{-1}A - \hat{\Lambda}' - \hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda}\hat{M}\hat{\Lambda}' = (\hat{M}^{-1} + \hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda})J'\hat{\Sigma},$$

where $j_{i\alpha} = 0$, $i \neq i(1, \alpha), \cdots, i(p_\alpha, \alpha)$, $\alpha = 1, \cdots, m$.

PROOF. The logarithm of the likelihood function of Ψ, given $\hat{\mu} = \bar{x}$, and divided by $N/2$ is

$$(10.4) \qquad \phi = p \log 2\pi - \log |\Psi| - \operatorname{tr} \Psi^{-1}A.$$

The partial derivative of ϕ with respect to ψ_{hg} (where ψ_{hg} is not assumed to be ψ_{gh}) is the g, hth element of

$$(10.5) \qquad \Psi^{-1}A\Psi^{-1} - \Psi^{-1}.$$

Then

$$\frac{\partial \phi}{\partial \sigma_{ii}} = \sum_{g,\,h} \frac{\partial \phi}{\partial \psi_{gh}} \frac{\partial \psi_{gh}}{\partial \sigma_{ii}} = \sum_{g',\,h'} \psi^{ig'} a_{g'h'} \psi^{h'i} - \psi^{ii}, \tag{10.6}$$

$$\frac{\partial \phi}{\partial m_{\alpha\beta}} = \sum_{g,\,h} \frac{\partial \phi}{\partial \psi_{gh}} \frac{\partial \psi_{gh}}{\partial m_{\alpha\beta}} = \sum_{g,\,g',\,h,\,h'} \lambda_{g\alpha} \psi^{gg'} a_{g'h'} \psi^{h'h} \lambda_{h\beta} - \sum_{g,\,h} \lambda_{g\alpha} \psi^{gh} \lambda_{h\beta}, \quad \alpha \neq \beta, \tag{10.7}$$

$$\frac{\partial \phi}{\partial \lambda_{i\alpha}} = \sum_{g,\,h} \frac{\partial \phi}{\partial \psi_{gh}} \frac{\partial \psi_{gh}}{\partial \lambda_{i\alpha}} = 2 \sum_{\beta,\,g} m_{\alpha\beta} \lambda_{g\beta} \left(\sum_{g',\,h'} \psi^{gg'} a_{g'h'} \psi^{h'i} - \psi^{gi} \right) \tag{10.8}$$

$$i \neq i\,(1,\,\alpha), \cdots, (p_\alpha,\,\alpha),$$

where $\psi^{-1} = (\Psi^{gh})$. Let

$$\hat{M}\hat{\Lambda}'(\hat{\Psi}^{-1}A\hat{\Psi}^{-1} - \hat{\Psi}^{-1}) = J'. \tag{10.9}$$

When we set (10.6), (10.7), and (10.8) equal to 0, we obtain

$$\operatorname{diag}(\hat{\Psi}^{-1}A\hat{\Psi}^{-1} - \hat{\Psi}^{-1}) = 0, \tag{10.10}$$

$$\hat{\Lambda}'(\hat{\Psi}^{-1}A\hat{\Psi}^{-1} - \hat{\Psi}^{-1})\hat{\Lambda} = D, \tag{10.11}$$

$$j_{i\alpha} = 0, \qquad i \neq i(1,\,\alpha), \cdots, i(p_\alpha,\,\alpha),\ \alpha = 1, \cdots, m, \tag{10.12}$$

where D is diagonal. The last set of equations states that J has 0's where Λ is not specified to have 0's.

Now let us simplify these equations; in particular, we want to express $\hat{\Psi}^{-1}A\hat{\Psi}^{-1} - \hat{\Psi}^{-1}$ differently. Multiplication of (10.11) on the left by $\hat{M}$ and (10.9) on the right by $\hat{\Lambda}$ gives

$$\hat{M}D = J'\hat{\Lambda}. \tag{10.13}$$

The diagonal elements of $J'\hat{\Lambda}$ are

$$\sum_i j_{i\alpha}\lambda_{i\alpha} = 0 \tag{10.14}$$

because either $j_{i\alpha} = 0$ or $\lambda_{i\alpha} = 0$. Thus the diagonal elements of $\hat{M}D$ are

$$\hat{m}_{\alpha\alpha}d_{\alpha\alpha} = d_{\alpha\alpha} = 0, \tag{10.15}$$

and therefore $D = 0 = J'\hat{\Lambda}$. Multiplication of (10.9) on the right by $\hat{\Psi}$ gives

$$\hat{M}\hat{\Lambda}'(\hat{\Psi}^{-1}A - I) = J'\hat{\Psi} = J'(\hat{\Sigma} + \hat{\Lambda}\hat{M}\hat{\Lambda}') = J'\hat{\Sigma} \tag{10.16}$$

because $J'\hat{\Lambda} = 0$. Then

$$\begin{aligned}(I + \hat{M}\hat{\Gamma})J'\hat{\Sigma} &= (I + \hat{M}\hat{\Lambda}'\hat{\Sigma}^{-1}\hat{\Lambda})\hat{M}\hat{\Lambda}'(\hat{\Psi}^{-1}A - I) \\ &= \hat{M}\hat{\Lambda}'\hat{\Sigma}^{-1}(\hat{\Sigma} + \hat{\Lambda}\hat{M}\hat{\Lambda}')(\hat{\Psi}^{-1}A - I) \\ &= \hat{M}\hat{\Lambda}'\hat{\Sigma}^{-1}(A - \hat{\Psi}).\end{aligned} \tag{10.17}$$

From this we derive (10.3). Now we can write

$$\begin{aligned}A - \hat{\Psi} &= \hat{\Psi}(\hat{\Psi}^{-1}A - I) \\ &= (\hat{\Sigma} + \hat{\Lambda}\hat{M}\hat{\Lambda}')(\hat{\Psi}^{-1}A - I) \\ &= \hat{\Sigma}(\hat{\Psi}^{-1}A - I) + \hat{\Lambda}J'\Sigma\end{aligned} \tag{10.18}$$

by (10.16). This can be written as

$$A - \hat{\Psi} = \hat{\Sigma}(\hat{\Psi}^{-1}A\hat{\Psi}^{-1} - \hat{\Psi}^{-1})\hat{\Psi} + \hat{\Lambda}J'\Sigma \tag{10.19}$$
$$= \hat{\Sigma}(\hat{\Psi}^{-1}A\hat{\Psi}^{-1} - \hat{\Psi}^{-1})\hat{\Sigma} + \hat{\Sigma}J\hat{\Lambda}' + \hat{\Lambda}J'\Sigma .$$

The diagonal elements of $\hat{\Sigma}J\hat{\Lambda}'$ and $\hat{\Lambda}J'\Sigma$ are 0 because $\sigma_{ii}\sum_{a}\lambda_{ia}j_{ia} = 0$ since j_{ia} is 0 if $\hat{\lambda}_{ia}$ is not 0. Then (10.19) implies (10.1).

We can find another set of equations for the estimates by eliminating J from (10.2) and (10.3). Multiplication of (10.3) on the right by $\hat{\Sigma}^{-1}\hat{\Lambda}$ gives

$$\hat{\Lambda}'\hat{\Sigma}^{-1}A\hat{\Sigma}^{-1}\hat{\Lambda} - \hat{\Gamma} - \hat{\Gamma}M\hat{\Gamma} = 0 \tag{10.20}$$

because $J'\hat{\Lambda} = 0$. Let $\hat{K} = \hat{\Lambda}'\hat{\Sigma}^{-1}A\hat{\Sigma}^{-1}\hat{\Lambda}$. Then

$$(I + \hat{M}\hat{\Gamma}) = \hat{\Gamma}^{-1}\hat{K} \tag{10.21}$$

and (10.17) gives

$$J'\hat{\Sigma} = \hat{K}^{-1}\hat{\Gamma}M\hat{\Lambda}'\hat{\Sigma}^{-1}(A - \hat{\Psi}) \tag{10.22}$$
$$= \hat{K}^{-1}\hat{\Gamma}M\hat{\Lambda}'\hat{\Sigma}^{-1}(A - \hat{\Sigma} - \hat{\Lambda}M\hat{\Lambda}')$$
$$= \hat{K}^{-1}\hat{\Gamma}\hat{M}(\hat{\Lambda}'\hat{\Sigma}^{-1}A - \hat{\Lambda}' - \hat{\Gamma}M\hat{\Lambda}')$$
$$= \hat{K}^{-1}\hat{\Gamma}\hat{M}(\hat{\Lambda}'\hat{\Sigma}^{-1}A - (I + \hat{\Gamma}\hat{M})\hat{\Lambda}')$$
$$= \hat{K}^{-1}\hat{\Gamma}\hat{M}(\hat{\Lambda}'\hat{\Sigma}^{-1}A - \hat{K}\hat{\Gamma}^{-1}\hat{\Lambda}') .$$

From (10.21) we also have $\hat{M}\hat{\Gamma} = \hat{\Gamma}^{-1}\hat{K} - I$, $\hat{K}^{-1}\hat{\Gamma}\hat{M} = \hat{\Gamma}^{-1} - \hat{K}^{-1}$, and (10.22) is

$$J'\hat{\Sigma} = (\hat{\Gamma}^{-1} - \hat{K}^{-1})(\hat{\Lambda}'\hat{\Sigma}^{-1}A - \hat{K}\hat{\Gamma}^{-1}\hat{\Lambda}') . \tag{10.23}$$

Then the estimates are defined by (10.1), (10.21) and the equations where the elements of (10.23) are set equal to zero if the corresponding element of Λ' is not assumed zero.

Let us consider a special case of $m = 2$. We order the rows of Λ so that

$$\Lambda = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}. \tag{10.24}$$

Then

$$J = \begin{pmatrix} 0 & c \\ d & 0 \end{pmatrix}. \tag{10.25}$$

We can write (10.2) as

$$J'\hat{\Lambda} = \begin{pmatrix} 0 & d'\hat{b} \\ c'\hat{a} & 0 \end{pmatrix} = 0 , \tag{10.26}$$

and (10.3) as

$$\begin{pmatrix} \hat{a}' & 0 \\ 0 & \hat{b}' \end{pmatrix}\hat{\Sigma}^{-1}A - \begin{pmatrix} \hat{a}' & 0 \\ 0 & \hat{b}' \end{pmatrix} - \begin{pmatrix} \hat{a}' & 0 \\ 0 & \hat{b}' \end{pmatrix}\hat{\Sigma}^{-1}\begin{pmatrix} \hat{a} & 0 \\ 0 & \hat{b} \end{pmatrix}\begin{pmatrix} 1 & \hat{m}_{12} \\ \hat{m}_{21} & 1 \end{pmatrix}\begin{pmatrix} \hat{a}' & 0 \\ 0 & \hat{b}' \end{pmatrix} \tag{10.27}$$
$$= \left[\hat{M}^{-1} + \begin{pmatrix} \hat{a}' & 0 \\ 0 & \hat{b}' \end{pmatrix}\hat{\Sigma}^{-1}\begin{pmatrix} \hat{a} & 0 \\ 0 & \hat{b} \end{pmatrix}\right]\begin{pmatrix} 0 & d' \\ c' & 0 \end{pmatrix}\hat{\Sigma}^{-1} .$$

If we let

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix}, \qquad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \tag{10.28}$$

then we have

$$(10.29)\quad \begin{pmatrix} \hat{a}'\hat{\Sigma}_1^{-1}A_{11} & \hat{a}'\hat{\Sigma}_1^{-1}A_{12} \\ \hat{b}'\hat{\Sigma}_2^{-1}A_{21} & \hat{b}'\hat{\Sigma}_2^{-1}A_{22} \end{pmatrix} - \begin{pmatrix} \hat{a}' & 0 \\ 0 & \hat{b}' \end{pmatrix} - \begin{pmatrix} \hat{a}'\hat{\Sigma}_1^{-1}\hat{a}\cdot\hat{a}' & \hat{a}'\hat{\Sigma}_1^{-1}\hat{a}\hat{m}_{12}\hat{b}' \\ \hat{b}'\hat{\Sigma}_2^{-1}\hat{b}\hat{m}_{21}\hat{a}' & \hat{b}'\hat{\Sigma}_2^{-1}\hat{b}\cdot\hat{b}' \end{pmatrix}$$
$$= \begin{pmatrix} -\dfrac{\hat{m}_{12}}{1-\hat{m}_{12}^2}\, c'\hat{\Sigma}_1^{-1} & \left(\dfrac{1}{1-\hat{m}_{12}^2} + \hat{a}'\hat{\Sigma}_1^{-1}\hat{a}\right) d'\hat{\Sigma}_2^{-1} \\ \left(\dfrac{1}{1-\hat{m}_{12}^2} + \hat{b}'\hat{\Sigma}_2^{-1}\hat{b}\right) c'\hat{\Sigma}_1^{-1} & -\dfrac{\hat{m}_{12}}{1-\hat{m}_{12}^2}\, \hat{d}'\hat{\Sigma}_2^{-1} \end{pmatrix}.$$

These can be written

$$(10.30)\quad \begin{aligned} \hat{a}'\hat{\Sigma}_1^{-1}A_{11} - \hat{a}' - (\hat{a}'\hat{\Sigma}_1^{-1}\hat{a})\,\hat{a}' &= -\frac{\hat{m}_{12}}{1-\hat{m}_{12}^2}\, c'\hat{\Sigma}_1^{-1}, \\ \hat{a}'\hat{\Sigma}_1^{-1}A_{12} - (\hat{a}_1'\hat{\Sigma}_1^{-1}\hat{a})\,\hat{m}_{12}\hat{b}' &= \left(\frac{1}{1-\hat{m}_{12}^2} + \hat{a}'\hat{\Sigma}_1^{-1}a\right) d'\hat{\Sigma}_2^{-1}, \\ \hat{b}'\hat{\Sigma}_2^{-1}A_{21} - \hat{b}'\hat{\Sigma}_2^{-1}\hat{b}\hat{m}_{21}\hat{a}' &= \left(\frac{1}{1-\hat{m}_{12}^2} + \hat{b}'\hat{\Sigma}_2^{-1}\hat{b}\right) c'\hat{\Sigma}_1^{-1}, \\ \hat{b}'\hat{\Sigma}_2^{-1}A_{22} - \hat{b}' - \hat{b}'\hat{\Sigma}_2^{-1}\hat{b}\hat{b}' &= -\frac{\hat{m}_{12}}{1-\hat{m}_{12}^2}\, d'\hat{\Sigma}_2^{-1}. \end{aligned}$$

11. Estimates for nonrandom factor scores when $\Lambda\Lambda'$ is unrestricted

Here

$$(11.1)\quad X_\alpha = \Lambda f_\alpha + U_\alpha + \mu, \qquad \alpha = 1, \cdots, N,$$

$$(11.2)\quad \mathcal{E}X_\alpha = \Lambda f_\alpha + \mu.$$

We assume

$$(11.3)\quad \sum_{\alpha=1}^{N} f_\alpha = 0,$$

$$(11.4)\quad \frac{1}{N-1}\sum_{\alpha=1}^{N} f_\alpha f_\alpha' = M.$$

Let

$$(11.5)\quad A = \frac{1}{N-1}\sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'.$$

In (11.4) and (11.5) we have altered previous definitions by replacing N by $N-1$. Then $(N-1)A$ has the noncentral Wishart distribution with covariance matrix Σ, means matrix $\Lambda M\Lambda'$, and $N-1=n$ degrees of freedom. Let $k_1, \cdots, k_m$ be the nonzero roots of

$$(11.6)\quad |\Lambda M\Lambda' - k\Sigma A^{-1}\Sigma| = 0.$$

Then [3] the likelihood function of A can be written[2]

$$(11.7)\quad C|\Sigma|^{-n/2}|A|^{(n-p-1)/2}e^{-n/2\,\mathrm{tr}\,\Sigma^{-1}A - n/2\,\mathrm{tr}\,M\Lambda'\Sigma^{-1}\Lambda}\int |I - ZZ'|^{(n-2m-1)/2}e^{n\,\mathrm{tr}\,K^{1/2}Z}dZ,$$

[2] In [3] the roots should have been defined by the equation $|T - \lambda\Sigma A^{-1}\Sigma| = 0$.

where the integration of the $m \times m$ elements of Z is over the range $I - ZZ'$ positive definite and

$$K^{1/2} = \begin{pmatrix} \sqrt{k_1} & 0 & \cdots & 0 \\ 0 & \sqrt{k_2} & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \sqrt{k_m} \end{pmatrix}. \tag{11.8}$$

We now take $M = I$. The indeterminacy remaining in Λ we remove for this particular sample by requiring that $\Lambda'\Sigma^{-1}A\Sigma^{-1}\Lambda$ be diagonal. Then

$$\hat{K} = \hat{\Lambda}'\hat{\Sigma}^{-1}A\hat{\Sigma}^{-1}\hat{\Lambda}\,. \tag{11.9}$$

If

$$\tfrac{1}{2}nF(K) = \log \int |I - ZZ'|^{(n-2m-1)/2} e^{n \operatorname{tr} K^{1/2}Z} dZ\,, \tag{11.10}$$

then

$$\frac{2}{n}\log L = \log C + \log|\Sigma^{-1}| + \frac{n-p-1}{n}\log|A| - \operatorname{tr}\Sigma^{-1}A - \operatorname{tr}\Lambda'\Sigma^{-1}\Lambda + F(K)\,. \tag{11.11}$$

The partial derivatives with respect to σ^{ii}, the elements of Σ^{-1}, and λ_{ia}, the elements of Λ, are

$$\frac{\partial \frac{2}{n}\log L}{\partial \sigma^{ii}} = \sigma_{ii} - a_{ii} - \sum_a \lambda_{ia}^2 + 2\sum_a F_{ka}(K) \sum_j \lambda_{ia}\lambda_{ja}\sigma^{ii}a_{ij} \tag{11.12}$$

$$\frac{\partial \frac{2}{n}\log L}{\partial \lambda_{ia}} = -2\sigma^{ii}\lambda_{ia} + 2F_{ka}(K) \sum_j \sigma^{ii}a_{ij}\sigma^{jj}\lambda_{ja}\,. \tag{11.13}$$

We set these derivatives equal to 0 to define the maximum likelihood estimates. These derivatives set equal to 0 give

$$\operatorname{diag}\hat{\Sigma} = \operatorname{diag}(A + \hat{\Lambda}\hat{\Lambda}' - 2\hat{\Lambda}F^*\hat{\Lambda}'\hat{\Sigma}^{-1}A), \tag{11.14}$$

$$\hat{\Sigma}^{-1}\hat{\Lambda} = \hat{\Sigma}^{-1}A\hat{\Sigma}^{-1}\hat{\Lambda}F^*\,, \tag{11.15}$$

where F^* is the diagonal matrix composed of the partial derivatives of $F(K)$. When we multiply (11.15) on the left by $\hat{\Lambda}'$, we obtain

$$\hat{\Gamma} = \hat{K}F^*\,, \tag{11.16}$$

which shows that $\hat{\Gamma}$ is diagonal. Then (11.15) is

$$\hat{\Lambda}\hat{\Gamma}^{-1}\hat{K} = A\hat{\Sigma}^{-1}\hat{\Lambda}\,. \tag{11.17}$$

From (11.14) and (11.15) we obtain

$$\operatorname{diag}\hat{\Sigma} = \operatorname{diag}(A + \hat{\Lambda}\hat{\Lambda}' - 2\hat{\Lambda}\hat{\Lambda}') = \operatorname{diag}(A - \hat{\Lambda}\hat{\Lambda}')\,. \tag{11.18}$$

The estimates are then defined by (11.9), (11.16), (11.17), (11.18), the requirement that $\hat{K}$ be diagonal and the definition of the diagonal elements of the diagonal matrix of F^*as the partial derivatives of (11.10) with respect to the diagonal elements of $K = \hat{K}$.

Equation (11.17) indicates that the diagonal elements of $(F^*)^{-1} = \hat{\Gamma}^{-1}\hat{K}$ must be m roots of

$$|A - \theta\hat{\Sigma}| = 0\,, \tag{11.19}$$

whereas in section 7.2 the m roots are the diagonal elements of $I + \hat{\Gamma}$. F^* is a complicated function of $\hat{K} = \hat{\Lambda}'\hat{\Sigma}^{-1}A\hat{\Sigma}^{-1}\hat{\Lambda}$.

We can show, however, that the estimates given in section 7.2 differ from the solution of the above equations (for a given A) by amounts smaller than order $1/\sqrt{n}$. To show this we make an asymptotic evaluation of F^*.

THEOREM 11.1. *For a given positive definite diagonal matrix H*

$$\lim_{n\to\infty} \sqrt{n}\left[\frac{\partial F(H)}{\partial h_a} - \frac{-1+\sqrt{1+4h_a}}{2h_a}\right] = 0 \tag{11.20}$$

uniformly in H for H in a bounded set, where h_a is the ath diagonal element of H and $F(H)$ is defined by (11.10).

The proof of this theorem is too complicated to give here. Using this result in (11.16) we have

$$\sqrt{n}\{\hat{\Gamma} - [-I + (I + 4\hat{K})^{1/2}]/2\} \to 0 \tag{11.21}$$

uniformly for $\hat{K}$ in a bounded set. Now if we replace (11.15) by

$$\hat{\Gamma} = [-I + (I + 4\hat{K})^{1/2}]/2 \tag{11.22}$$

the solution of the equations will differ from the previous solution by amounts smaller than order $1/\sqrt{n}$ (because the solutions are continuous). However, (11.22) is equivalent to $\hat{K} = \hat{\Gamma}(I + \hat{\Gamma})$ and then the solution is the one of section 7.2. The errors are uniformly of order $o(1/\sqrt{n})$ for A in a bounded set and for large enough n the probability that A is in a set including Ψ is arbitrarily near one. Hence

THEOREM 11.2. *If A converges stochastically to a positive definite matrix, then $\sqrt{n}$ times the difference between the estimates of $\hat{\Lambda}$ and $\hat{\Sigma}$ defined in this section and the estimates defined in section* 7.2 *converge stochastically to* 0.

12. Asymptotic normality of estimates

In this section we shall prove that $\sqrt{N}(\hat{\Lambda} - \Lambda)$ and $\sqrt{N}(\hat{\Sigma} - \Sigma)$ defined by (7.4) to to (7.7) are asymptotically normally distributed when A converges stochastically to $\Psi = \Sigma + \Lambda\Lambda'$ and $\sqrt{N}(A - \Psi)$ is asymptotically normally distributed. In particular, this is true when the observations $x_1, x_2, \cdots$ are drawn from $N(\mu, \Psi)$. Let

$$\Phi = \Sigma - \Lambda(\Lambda'\Sigma^{-1}\Lambda)^{-1}\Lambda' . \tag{12.1}$$

THEOREM 12.1. *If $|\phi_{ij}^2| \neq 0$, where (ϕ_{ij}) is defined by* (12.1), *if Λ and Σ are identified by the condition that $\Lambda'\Sigma^{-1}\Lambda$ is diagonal and the diagonal elements are different and ordered, if A converges stochastically to Ψ and if $\sqrt{N}(A - \Psi)$ has a limiting normal distribution, then $\sqrt{N}(\hat{\Lambda} - \Lambda)$, $\sqrt{N}(\hat{\Sigma} - \Sigma)$ defined by* (7.4) *to* (7.7) *have a limiting normal distribution.*

PROOF: First we show that $\hat{\Lambda}$ converges stochastically to Λ and $\hat{\Sigma}$ converges stochastically to Σ. The estimates $\hat{\Lambda}$, $\hat{\Sigma}$ are defined as the matrices satisfying $\Lambda^{*\prime}\Sigma^{*-1}\Lambda^*$ being diagonal and maximizing

$$f(A, \Sigma^*, \Lambda^*) = \frac{2}{N}\log L(A, \Sigma^*, \Lambda^*) = -p\log 2\pi - \log|\Sigma^* + \Lambda^*\Lambda^*| - \operatorname{tr} A(\Sigma^* + \Lambda^*\Lambda^*)^{-1} . \tag{12.2}$$

Now

$$f(A, \Sigma^*, \Lambda^*) \to f(\Lambda\Lambda' + \Sigma, \Sigma^*\Lambda^*) \tag{12.3}$$

uniformly in probability in a neighborhood of Σ, Λ and $f(\Lambda\Lambda' + \Sigma, \Sigma^*, \Lambda^*)$ has a unique maximum at $\Sigma^* = \Sigma$, $\Lambda^* = \Lambda$. Because the functions are continuous, the Σ^* Λ^* maximizing $f(A, \Sigma^*, \Lambda^*)$ must converge stochastically to Σ, Λ.

To prove the theorem we need only to prove that $\hat{\Lambda}$ and $\hat{\Sigma}$ are functions of A that have continuous first derivatives in a neighborhood of $A = \Psi$. Since the equations defining $\hat{\Lambda}$ and $\hat{\Sigma}$ are rational functions of $\hat{\Lambda}$, $\hat{\Sigma}$ and A set equal to zero, they are polynomial equations; the derivatives will be continuous unless they become infinite. The remainder of the proof is to show that they do not become infinite (at $A = \Psi$).

$\hat{\Lambda}$ and $\hat{\Sigma}$ are defined implicitly as functions of A, say by

$$H(\hat{\lambda}, \hat{\sigma}, a) = 0\,, \tag{12.4}$$

where $\hat{\lambda}$ is $\hat{\Lambda}$ arranged in a vector, $\hat{\sigma}$ is $\hat{\Sigma}$ arranged in a vector and a is A arranged in a vector.

The solution to (12.4) is $\hat{\lambda} = \hat{\lambda}(a)$, $\hat{\sigma} = \hat{\sigma}(a)$. Then

$$H_{\hat{\lambda}}[\hat{\lambda}(a), \hat{\sigma}(a), a]\hat{\lambda}_a(a) + H_{\hat{\sigma}}[\hat{\lambda}(a), \hat{\sigma}(a), a]\hat{\sigma}_a(a) + H_a[\hat{\lambda}(a), \hat{\sigma}(a), a] \equiv 0\,, \tag{12.5}$$

where $H_{\hat{\lambda}}$ is the matrix of partial derivatives of the components of $H(\hat{\lambda}, \hat{\sigma}, a)$ with respect to the components of $\hat{\lambda}$, $\hat{\lambda}_a(a)$ is the matrix of partial derivatives of the components of $\hat{\lambda}(a)$ with respect to the components of a, etc. We need to show that

$$H_{\hat{\lambda}}(\lambda, \sigma, \psi)\hat{\lambda}_a(\psi) + H_{\hat{\sigma}}(\lambda, \sigma, \psi)\hat{\sigma}_a(\psi) + H_a(\lambda, \sigma, \psi) = 0 \tag{12.6}$$

can be solved for $\hat{\lambda}_a(\psi)$, $\hat{\sigma}_a(\psi)$. Our method of computation is to expand $H(\hat{\lambda}, \hat{\sigma}, a) = H(\lambda + l, \sigma + s, \psi + a^*)$ in terms of l, s, and a^*, consider only linear terms, and show (under the conditions of the theorem) that the resulting linear-equations can be solved for l and s in terms of a.

Let $\hat{\Lambda} = \Lambda + L$, $\hat{\Sigma} = \Sigma + S$, $A = \Sigma + \Lambda\Lambda' + A^*$, $\hat{\Gamma} = \Gamma + G$. Then (7.5) can be written in linear terms as

$$\operatorname{diag} S = \operatorname{diag}(A^* - \Lambda L' - L\Lambda')\,. \tag{12.7}$$

Since $\hat{\Sigma}^{-1} = (\Sigma + S)^{-1}$ is $\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}$ to linear terms in S, (7.6) can be written (in linear terms) as

$$\Gamma + G = \Gamma + L'\Sigma^{-1}\Lambda + \Lambda'\Sigma^{-1}L - \Lambda'\Sigma^{-1}S\Sigma^{-1}\Lambda\,, \tag{12.8}$$

and (7.8) can be written (in linear terms) as

$$\begin{aligned}\Lambda\Gamma + \Lambda L'\Sigma^{-1}\Lambda + \Lambda\Lambda'\Sigma^{-1}L - \Lambda\Lambda'\Sigma^{-1}S\Sigma^{-1}\Lambda + L\Gamma \\ = \Lambda\Lambda'\Sigma^{-1}\Lambda + \Lambda\Lambda'\Sigma^{-1}L - \Lambda\Lambda'\Sigma^{-1}S\Sigma^{-1}\Lambda + A^*\Sigma^{-1}\Lambda - S\Sigma^{-1}\Lambda\,.\end{aligned} \tag{12.9}$$

Then (12.9) and (7.7) can be written (in linear terms) as

$$L\Gamma + \Lambda L'\Sigma^{-1}\Lambda + S\Sigma^{-1}\Lambda = A^*\Sigma^{-1}\Lambda\,, \tag{12.10}$$

$$\operatorname{nondiag}(L'\Sigma^{-1}\Lambda + \Lambda'\Sigma^{-1}L - \Lambda'\Sigma^{-1}S\Sigma^{-1}\Lambda) = \operatorname{nondiag} 0\,. \tag{12.11}$$

We now show that (12.7), (12.10) and (12.11) can be solved for S and L under the conditions of the theorem. From (12.10) we derive

$$L = A^*\Sigma^{-1}\Lambda\Gamma^{-1} - S\Sigma^{-1}\Lambda\Gamma^{-1} - \Lambda L'\Sigma^{-1}\Lambda\Gamma^{-1} \tag{12.12}$$

and

$$L\Lambda' = A^*\Sigma^{-1}H - S\Sigma^{-1}H - \Lambda L'\Sigma^{-1}H, \tag{12.13}$$

where $H = \Lambda\Gamma^{-1}\Lambda'$. Using (12.13) we have

$$\begin{aligned} A^* - L\Lambda' - \Lambda L' &= A^* - A^*\Sigma^{-1}H + S\Sigma^{-1}H + \Lambda L'\Sigma^{-1}H \\ &\quad - H\Sigma^{-1}A^* + H\Sigma^{-1}S + H\Sigma^{-1}L\Lambda' \\ &= A^* - A^*\Sigma^{-1}H + S\Sigma^{-1}H + \Lambda L'\Sigma^{-1}H \\ &\quad - H\Sigma^{-1}A^* + H\Sigma^{-1}S + H\Sigma^{-1}(A^*\Sigma^{-1}H \\ &\quad - S\Sigma^{-1}H - \Lambda L'\Sigma^{-1}H). \end{aligned} \tag{12.14}$$

Since $H\Sigma^{-1}\Lambda = \Lambda$, (12.14) can be combined with (12.7) to give

$$\text{diag}\,[(\Sigma - H)\Sigma^{-1}S\Sigma^{-1}(\Sigma - H)] = \text{diag}\,[(\Sigma - H)\Sigma^{-1}A^*\Sigma^{-1}(\Sigma - H)] \tag{12.15}$$

or

$$\text{diag}\,\Phi\Sigma^{-1}S\Sigma^{-1}\Phi = \text{diag}\,\Phi\Sigma^{-1}A^*\Sigma^{-1}\Phi. \tag{12.16}$$

The ith component equation is

$$\sum_{j=1}^{p} \phi_{ij}^2 \frac{s_{jj}}{\sigma_{jj}^2} = \sum_{g,\,h=1}^{p} \phi_{ig} \frac{a_{gh}^*}{\sigma_{gg}\sigma_{hh}} \phi_{ih}. \tag{12.17}$$

This can be solved if the matrix Ξ with elements $\xi_{ij} = \phi_{ij}^2$ is nonsingular.

Equation (12.11) can be written as

$$\text{nondiag}\,(Q + Q') = \text{nondiag}\,V, \tag{12.18}$$

where $Q = L'\Sigma^{-1}\Lambda$ and $V = \Lambda'\Sigma^{-1}S\Sigma^{-1}\Lambda$. Multiplication of (12.10) on the left by $\Lambda'\Sigma^{-1}$ gives

$$Q'\Gamma + \Gamma Q = U - V, \tag{12.19}$$

where $U = \Lambda'\Sigma^{-1}A^*\Sigma^{-1}\Lambda$. The α, αth equation of (12.19) yields

$$q_{\alpha\alpha} = \frac{u_{\alpha\alpha} - v_{\alpha\alpha}}{2\gamma_{\alpha\alpha}}, \tag{12.20}$$

and the α, βth equations of (12.18) and (12.19) yield

$$q_{\alpha\beta} = \frac{u_{\alpha\beta} - v_{\alpha\beta} - v_{\alpha\beta}\gamma_{\beta\beta}}{\gamma_{\alpha\alpha} - \gamma_{\beta\beta}}, \qquad \alpha \neq \beta. \tag{12.21}$$

Substitution of $Q = L'\Sigma^{-1}\Lambda$ in (12.12) gives L in terms of A^* and S. This proves the theorem.

From the above formulas we can find the asymptotic variances and covariances of the estimates. When the observations are drawn from a normal distribution with covariance matrix Ψ,

$$\begin{aligned} \frac{N^2}{N-1}\,\mathcal{E}\,(a_{gh} - \psi_{gh})\,(a_{kl} - \psi_{kl}) &= \frac{N^2}{N-1}\,\mathcal{E}\,a_{gh}^* a_{kl}^* \\ &= \psi_{gk}\psi_{hl} + \psi_{gl}\psi_{hk}. \end{aligned} \tag{12.22}$$

and this is the limiting covariance of A^*_{gh} and A^*_{kl}. The solution of (12.17) can be expressed as

$$\frac{s_{kk}}{\sigma^2_{kk}} = \sum_{i,\,g,\,h} \xi^{ki}\phi_{ig}\frac{a^*_{gh}}{\sigma_{gg}\sigma_{hh}}\phi_{ih}, \tag{12.23}$$

where $(\xi^{ki}) = \Xi^{-1}$. Then it is a straightforward computation to find that the limiting covariance of s_{kk}/σ^2_{kk} and s_{gg}/σ^2_{gg} is 2 ξ^{kg}. This shows that

$$\lim_{N\to\infty} N\mathcal{E}\,(\hat{\sigma}_{kk} - \sigma_{kk})\,(\hat{\sigma}_{gg} - \sigma_{gg}) = 2\xi^{kg}\sigma^2_{gg}\sigma^2_{gg}. \tag{12.24}$$

The limiting covariances of elements of $\hat{\Lambda}$ are complicated and depend on the identification conditions. We give one formula to show how involved such expressions are and to show that they are similar to formulas obtained in other problems involving characteristic vectors ([16], for example). Let λ_ν be the νth column of Λ and $\hat{\lambda}_\nu$ the νth column of $\hat{\Lambda}$. Then

$$\gamma_{\nu\nu}\lim_{N\to\infty} N\mathcal{E}\,(\lambda_\nu - \lambda_\nu)\,(\hat{\lambda}_\nu - \lambda_\nu)' = M_\nu\Sigma^{-1}\,(\Psi + 4\lambda_\nu\lambda_\nu')\,\Sigma^{-1}M_\nu + P_\nu\,(\lambda_{k\nu}\xi^{kg}\lambda_{g\nu})\,P_\nu, \tag{12.25}$$

where $(\lambda_{k\nu}\xi^{kg}\lambda_{g\nu})$ is a matrix with indicated elements and

$$M_\nu = \Sigma - \frac{1}{2\gamma_{\nu\nu}}\lambda_\nu\lambda_\nu' - \sum_{\alpha\neq\nu}\frac{1}{\gamma_{\alpha\alpha} - \gamma_{\nu\nu}}\lambda_\alpha\lambda_\alpha', \tag{12.26}$$

$$P_\nu = \Sigma - \frac{1}{2\gamma_{\nu\nu}} - (1 + \gamma_{\nu\nu})\sum_{\alpha\neq\nu}\frac{1}{\gamma_{\alpha\alpha} - \gamma_{\nu\nu}}\lambda_\alpha\lambda_\alpha'. \tag{12.27}$$

The method of estimation of section 7.4 (principal components of $A - \hat{\Sigma}$) can be studied in a similar fashion and one can derive the following result:

THEOREM 12.2. *If* $|\theta^2_{ij}| \neq 0$ *where* $\Theta = I - \Lambda(\Lambda'\Lambda)^{-1}\Lambda'$, *if* Λ *is identified by the condition that* $\Lambda'\Lambda$ *is diagonal and the diagonal elements are different and ordered, if* A *converges stochastically to* Ψ, *and if* $\sqrt{N}(A - \Psi)$ *has a limiting normal distribution, then* $\sqrt{N}(\hat{\Lambda} - \Lambda)$, $\sqrt{N}(\hat{\Sigma} - \Sigma)$, *defined by* (7.22) *to* (7.25), *has a limiting normal distribution.*

We can also prove that when Λ is identified by requiring certain elements to be 0 then the estimates are also asymptotically normally distributed. Instead of requiring $m_{ii} = 1$, we can normalize each column of Λ by a restriction on that column (for example, requiring $\lambda_{\nu\nu} = 1$, $\nu = 1, \cdots, m$, if none of these is specified to be 0). Then all of the restrictions are on Λ. This is desirable because then we can compare populations that differ only in M (see section 7.9) and we can compare the case of random and nonrandom factor scores. We can then prove a striking and powerful theorem. Let

$$M(N) = \frac{1}{N}\sum_{\alpha=1}^{N}[f_\alpha - \bar{f}(N)][f_\alpha - \bar{f}(N)]', \tag{12.28}$$

$$\sigma_{ii}(N) = \frac{1}{N}\sum_{\alpha=1}^{N}[u_{i\alpha} - \bar{u}_i(N)]^2, \tag{12.29}$$

where

$$\bar{f}(N) = \frac{1}{N}\sum_{\alpha=1}^{N} f_\alpha, \qquad \bar{u}_i(N) = \frac{1}{N}\sum_{\alpha=1}^{N} u_{i\alpha}, \tag{12.30}$$

and

$$(12.31)\qquad b_{ij}(N) = \frac{1}{\sqrt{N}}\sum_{a=1}^{N}[u_{ia} - \bar{u}_i(N)][u_{ja} - \bar{u}_j(N)], \qquad i \neq j,$$

and let $\Sigma(N)$ be the diagonal matrix with (12.29) as elements.

THEOREM 12.3. *If Λ is identified by specified zero elements, if the identification and normalization is by restrictions on Λ, if $M(N)$ and $\Sigma(N)$ approach limits in probability and if $b_{ij}(N)$ have a limiting joint normal distribution with zero means, then $\sqrt{N}(\hat{\Lambda} - \Lambda)$, $\sqrt{N}[\hat{M} - M(N)]$ and $\sqrt{N}[\hat{\Sigma} - \Sigma(N)]$ are asymptotically normally distributed, where $\hat{\Lambda}$, $\hat{M}$, $\hat{\Sigma}$ are the maximum likelihood estimates of section 7.6 normalized by the restrictions on $\hat{\Lambda}$. If*

$$(12.32)\qquad \lim_{N\to\infty} \mathcal{E} b_{ij}(N)\, b_{kl}(N) = \operatorname*{plim}_{N\to\infty} \sigma_{ii}(N)\,\sigma_{jj}(N), \qquad i = k,\ j = l$$
$$i = l,\ j = k,$$
$$= 0, \qquad \textit{otherwise},$$

then the parameters of the limiting normal distribution of the estimates depend only on Λ, plim $M(N)$ and plim $\Sigma(N)$.

The proof of this theorem is too involved to give here. It should be noted that if f and U are normally distributed, the theorem holds.

REFERENCES

[1] A. A. ALBERT, "The matrices of factor analysis," *Proc. Nat. Acad. Sci.*, Vol. 30 (1944), pp. 90–95.

[2] ———, "The minimum rank of a correlation matrix," *Proc. Nat. Acad. Sci.*, Vol. 30 (1944), pp. 144–146.

[3] T. W. ANDERSON, "The non-central Wishart distribution and certain problems of multivariate statistics," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 409–431.

[4] ———, "Asymptotic theory for principal component analysis," to be published.

[5] M. S. BARTLETT, "The statistical conception of mental factors," *Brit. Jour. Psych.*, Vol. 28 (1937), pp. 97–104.

[6] ———, "Tests of significance in factor analysis," *Brit. Jour. Psych.* (Stat. Sec.), Vol. 3 (1950), pp. 77–85.

[7] ———, "A further note on tests of significance in factor analysis," *Brit. Jour. Psych.* (Stat. Sec.), Vol. 4 (1951), pp. 1–2.

[8] ———, "The effect of standardization on a χ^2 approximation in factor analysis," *Biometrika*, Vol. 38 (1951), pp. 337–344.

[9] ———, "Factor analysis in psychology as a statistician sees it," *Uppsala Symposium on Psychological Factor Analysis, 17–19 March 1953*, Uppsala, Almqvist and Wiksell, 1953, pp. 23–34.

[10] K. J. HOLZINGER and H. H. HARMON, *Factor Analysis*, Chicago, University of Chicago Press, 1941.

[11] HAROLD HOTELLING, "Analysis of a complex of statistical variables into principal components," *Jour. Educational Psych.*, Vol. 24 (1933), pp. 417–441, 498–520.

[12] M. G. KENDALL and B. BABINGTON SMITH, "Factor analysis," *Jour. Roy. Stat. Soc.*, Ser. B, Vol. 12 (1950), pp. 60–94.

[13] T. C. KOOPMANS and O. REIERSØL, "The identification of structural characteristics," *Annals of Math. Stat.*, Vol. 21 (1950), pp. 165–181.

[14] D. N. LAWLEY, "The estimation of factor loadings by the method of maximum likelihood," *Proc. Roy. Soc. Edin.*, Vol. 60 (1940), pp. 64–82.

[15] ———, "Further investigations in factor estimation," *Proc. Roy. Soc. Edin.*, Vol. 61 (1942), pp. 176–185.

[16] ———, "A modified method of estimation in factor analysis and some large sample results," *Uppsala Symposium on Psychological Factor Analysis, 17–19 March 1953*, Uppsala, Almqvist and Wiksell, 1953, pp. 35–42.

[17] C. R. RAO, "Estimation and tests of significance in factor analysis," *Psychometrika*, Vol. 20 (1955), pp. 93–111.

[18] G. RASCH, "On simultaneous factor analysis in several populations," *Uppsala Symposium on Psychological Factor Analysis, 17–19 March 1953*, Uppsala, Almqvist and Wiksell, 1953, pp. 65–71.

[19] OLAV REIERSØL, "On the identifiability of parameters in Thurstone's multiple factor analysis," *Psychometrika*, Vol. 15 (1950), pp. 121–149.

[20] CHARLES SPEARMAN, "General intelligence, objectively determined and measured," *Amer. Jour. Psych.*, Vol. 15 (1904), pp. 201–293.

[21] G. H. THOMSON, "Hotelling's method modified to give Spearman's *g*," *Jour. Educational Psych.*, Vol. 25 (1934), pp. 366–374.

[22] ———, "Some points of mathematical technique in the factorial analysis of ability," *Jour. Educational Psych.*, Vol. 27 (1936), pp. 37–54.

[23] ———, *The Factorial Analysis of Human Ability*, 5th ed., London, University of London Press, 1953

[24] L. L. THURSTONE, *Multiple-factor Analysis*, Chicago, University of Chicago Press, 1947.

[25] E. B. WILSON, "Review of *Crossroads in the Mind of Man* (by T. L. Kelley)," *Jour. General Psych.*, Vol. 2 (1929), pp. 153–169.

[26] E. B. WILSON and JANE WORCESTER, "The resolution of tests into two general factors," *Proc. Nat. Acad. Sci.*, Vol. 25 (1939), pp. 20–25.

[27] ———, "The resolution of six tests into three general factors," *Proc. Nat. Acad. Sci.*, Vol. 25 (1939), pp. 73–77.

[28] P. WHITTLE, "On principal components and least square methods of factor analysis," *Skand. Aktuar.*, Vol. 35 (1952), pp. 223–239.